Brain Organization and Information Integration

By

Erik Hoel

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Neuroscience)

at the

UNIVERSITY OF WISCONSIN-MADISON

2016

Date of final oral examination: 1/20/16

This dissertation is approved by the following members of the Final Oral Committee:

Giulio Tononi, Professor, Psychiatry
Chiara Cirelli, Professor, Psychiatry
Bradley R. Postle, Professor, Psychology
Barry D. Van Veen, Professor, Engineering
Larry Shapiro, Professor, Philosophy

Dedication

I dedicate this thesis to Julia Buntaine, who has supported me graciously and lovingly from both near and afar during my time in Wisconsin.

Acknowledgements

The work detailed in this thesis would not have been possible without the help of a great many people, especially from my friends, co-workers, and mentors.

First, I would like to thank my colleagues in the NTP. They have been my classmates and my friends, and it's with them that I've gone through the peaks and valleys of graduate school life. Special thanks to fellow NTPers Chadd Funk and Alex Rodriguez, who, on top of being my friends, have also been great co-workers as well.

Second, I thank the colleagues with whom I have worked in the "Skynet" section of the lab, such as Puneet Rana, Andy Nere, and Will Mayner. Thanks to Umberto Olcese for answering many questions about Synthesis via email. Thanks to Billy Marshall for being a wonderful intellectual tennis partner – bouncing the ball of an idea back and forth has lead us to some of the formalizations contained herein. Extra special thanks to Larissa Albantakis, whose patience, fortitude, and intelligence have made her invaluable to the research I have performed over the years.

Third, I thank the mentors who have provided me with sagacious advice. My committee members, Brad Postle, Barry Van Veen, and Larry Shapiro, have all been instrumental in encouraging me while at the same time pointing out where there is room for improvement. Chiara Cirelli deserves so much thanks for being a great exemplar of a sharp-minded scientist. Her commitment has been inspirational. Giulio Tononi – there is far too much to say. But let me, at a minimum, thank Giulio for teaching me to how to think – to *really really* think. In my five years here I have been privileged to participate in and watch the development of an incredibly advanced technical and philosophical

achievement. The lessons I learned in helping erect such grand edifices will stay with me for the rest of my life, in all my endeavors.

Table of Contents

Introduction **1**

Chapter 1 **17**

Quantifying causal emergence shows that macro can beat micro

Chapter 2 **45**

Can the macro beat the micro? Integrated Information across spatiotemporal scales

Chapter 3 **80**

Synaptic refinement during development and its effect on slow wave activity – a computational study

Discussion **126**

Appendix I: Chapter 1 **135**

Appendix II: Chapter 2 **149**

Introduction

The research program of the thesis

My interest in the relationship between brain connectivity and consciousness dates back to my time as an undergraduate. From 2008-2010 I worked as the manager of an EEG lab, using graph theory to measure functional connectivity in EEG data. My senior thesis investigated changes in the "small-worldness" of whole brain functional connectivity during induced relative blindsight, wherein participants had changes in consciousness without changes in performance (Hoel and Hogan, 2010). My focus was on the balance of local vs. global interactions over the entire cortical functional network, and at the time there had been little interest in characterizing the neural correlates of consciousness with respect to the underlying structure of the network. I was drawn to the Center for Sleep and Consciousness because the Integrated Information Theory of Consciousness (IIT) offered a possible way to relate consciousness to network organization. My interest grew from there, and my thesis work has converged on two related inquiries:

I) How does the capacity to integrate information vary across spatiotemporal scales, and at which level does it reach a maximum?

II) How does the capacity to integrate information relate to the basic organization of the cortex? And how might it change with synaptic refinement during neural development?

My thesis consists of three related Aims: #1 and #2 establish the theoretical and computational tools needed to evaluate information integration at different spatiotemporal scales. Aim #3 uses large-scale simulations to show that synaptic refinement leads to characteristic changes in the activity patterns of thalamocortical networks, such as a marked decrease in slow wave activity during sleep. However, the

ultimate goal of this thesis, like all of science, is to set the stage for future work. Together these three Aims provide a foundation for the two projects that are detailed in the Future Directions. One of these projects is well underway in the lab, while the other is soon to be underway in association with the Human Brain Project.

Earlier, the results of my thesis work were instrumental in preparing a grant proposal on the fundamental nature of information and causality (with a project team of Giulio Tononi as PI, Christof Koch and Christoph Adami as project consultants), which was funded by Templeton World Charity Foundation in 2013.

Below, the background of the thesis is laid out, accompanied by an overview of the respective Aims themselves.

Neuroscience and the spatiotemporal scale of causal interactions

Neuroscientists observe brain activity across a multitude of different spatial and temporal scales using tools such as EEG, fMRI, local field potentials or single cell recordings. Hence, processing and representation of sensory stimuli, storage and retrieval of memories, and decision-making can be investigated at the scale of cortical regions, columns, minicolumns, individual neurons, and even individual synapses and ionic channels. How should these different spatial scales be related to one another? And are they all equally important, or are some levels privileged with respect to causal explanation? The same problem exists in the temporal domain of brain activity measurement (Marom, 2010): should one measure the interactions of brain regions over tens of milliseconds, over several hundred milliseconds, or even over the longer timescales of plastic changes (seconds, minutes, days, etc)? Take for instance the search

for the neural code, a central concern of modern neuroscience (Kumar et al., 2010). The search is typically addressed by relying on an information-theoretic account of brain activity. But what should one make of the fact that the spatiotemporal scale of measurement chosen by the observer drastically affects whether an event is information rich or barren (Panzeri et al., 2010)?

The question of spatiotemporal scale is particularly relevant in view of massive projects that are underway in present-day neuroscience, such as the Human Connectome Project (Sporns et al., 2005), the Blue Brain Project (Markram, 2006), along with the more recent Brain Activity Map and the Human Brain Project (Kandel et al., 2013). Some of these projects seek to map the structure and activity of the brain down to the synaptic level. Clearly, some perspective on how, why, and at which level to construct models of brain function is needed. For instance, a model of the brain can be constructed at the macro scale of regions and pathways, a meso scale of local populations of neurons such as minicolumns and their inter and intraconnectivity, and a micro scale of neurons and their synapses (Sporns et al., 2005). Are all these levels equally important, or are some hopelessly coarse-grained and others uselessly fine-grained? Answering this question is crucial, since wiring diagrams, even of the finest detail, do not guarantee causal explanation.

High-level versus low-level explanations have been in tension throughout neuroscience, and continue to be today. Beyond the long-running modularity vs. holism debate in neuroscience, consider more recently the evidence for both population-level coding (Georgopoulos et al., 1986) and sparse coding in cells colloquially referred to as "grandmother cells" (Quiroga et al., 2008). Or consider proposals that, rather than the

level of individual synapses or single neurons, it is the cortical minicolumn that is the primary functional unit of brain organization (Buxhoeveden and Casanova, 2002). In terms of temporal scale, consider the debate between burst coding and sparse coding: highly efficient temporal neural codes have been "discovered" in which each single spike contains information for an observer (Borst and Theunissen, 1999). This has been taken as evidence for a "sparse" neural code, in which the individual spikes and their timing carry information. Yet recent studies of the rat barrel cortex have shown a large amount of intrinsic noise in the cortex, as well as high average firing rates which are non-sparse, suggesting that *in vivo* neuronal networks may be chaotic and thus require a robust neural code (bursting) to maintain signal (London et al., 2010; Vijayan et al., 2010). This tension even extends, for example, to high-vs.-low types of explanations for the effects of anesthesia (Mashour, 2014).

Integrated information and the spatiotemporal grain of maximal cause-effect power

A possible approach to the question of scale is offered by Integrated Information Theory (IIT; Oizumi, Albantakis, & Tononi, 2014; G Tononi, 2004, 2008, 2012a, 2012b). Taking the *intrinsic perspective* of the brain, rather than the *extrinsic perspective* of an external observer, IIT asks: *what kind of neural events matter to the brain itself*?

According to IIT, the information contained in a neural event is not how "surprised" the observing scientist is, but rather based on the causal relationships among the parts of the neural system itself. Starting from the notion that what is important in a system are the "differences that make a difference" (Bateson, 1972), IIT seeks to characterize information from the intrinsic perspective of a system of elements

(mechanisms in a particular state). Integrated information (Φ) assesses the degree to which the system is irreducible to independent parts, over the minimum information partition (MIP). Φ is used to identify complexes – sets of elements that are maximally irreducible causally – i.e. they are endowed with maximal cause-effect power (Φ^{Max}).

Previous to the work in this thesis, integrated information had been characterized only over Markovian systems defined at a single spatial and temporal scale – a micro level of description over "atomic" spatial elements and elementary temporal intervals. However, IIT posits that in any given system of mechanisms there is going to be a particular spatiotemporal scale at which Φ reaches a maximum: the scale at which a system "self-defines" by having maximum cause-effect power. Moreover, the spatiotemporal scale at which Φ reaches a maximum is not necessarily the micro level (Tononi, 2008). Thus, Φ should be assessed at different spatiotemporal grains. Importantly, the particular spatiotemporal grain at which Φ is maximal in the brain is an open question: specifically, is the spatial grain of maximal cause-effect power that of neurons, neuronal ensembles, minicolumns, or even larger, such as over entire regions? Similarly, is the optimal temporal grain that of milliseconds, hundreds of milliseconds, or seconds or longer? (Albantakis et al., 2012).

Aim #1

To address these issues, we carried out a research program with the goal of establishing that Φ can indeed peak at a macro spatiotemporal scale. Aim #1's goal was to quantify causal interactions at multiple spatiotemporal grains in discrete systems, and show that causal measures (*effective information*) can peak at a macro scale. This contrasts with the

common assumption that, although macro descriptions may be convenient, only the micro level of a physical system is causally complete, because it includes every detail, thus leaving no room for causation at the macro level (a reductionist argument formalized by Kim (2000)). By developing an explicit analysis of causal relationships at different spatiotemporal scales we were able to show, in a principled manner, that causal interactions can be maximal at a macro scale – in other words, that there can be true causal emergence.

The work from Aim #1 was published as the paper "Quantifying causal emergence shows that macro can beat micro" in *The Proceedings of the National Academy of Sciences* and is included in full as Chapter 1. In the Aim we made use of simple systems, including neural-like ones, to show quantitatively that macro causal models can be superior to their underlying micro causal models. For each system, we completely characterized the causal mechanisms at the micro level, fixing what can happen at any supervening macro scale. Macro levels were defined by coarse graining the micro elements in space and/or time, and this mapping defined the repertoire of possible causes and effects at each level. Causation was measured in a state-dependent way via *cause-effect information*, and also in a state-independent manner with *effective information* (Tononi and Sporns, 2003). After comparing causal relationships at different levels, we showed that, depending on how a system is organized, *cause-effect information* and *effective information* could peak at a macro rather than at a micro level. This was shown by coarse-graining spatially, temporally, and spatiotemporally.

Overall, this paper accomplished Aim #1 by showing that it is in-principle possible that macro causal models can have greater causal power than micro causal

models (despite the macro supervening on the micro). Thus, the macro may be causally superior to the micro even though it supervenes upon it. Evaluating the changes in causation that arise from coarse or fine graining a system provided a straightforward way of quantifying both emergence and reduction. Additionally, it provided us with knowledge about what systems and network organizational characteristics lead to causal emergence – such as size of the system's state space, noise (causal divergence) and degeneracy (causal convergence).

Aim #2

The goal of Aim #2 was to extend Aim #1 by finding the spatiotemporal grain at which causal interactions are *maximally irreducible*, i.e., which grain constitutes a maximum of *integrated information* (Φ^{max}). The possibility of causal emergence was established in Aim #1 (Hoel et al., 2013) using *effective information*. However, IIT requires that, from the intrinsic perspective of a system, causal interactions are also *maximally irreducible* (Tononi, 2012a). Moreover, IIT also allows for the *composition* of causal interactions within a system. That is, IIT is sensitive to the possibility that multiple subsets of a system mediate distinct causal interactions, as long as they are irreducible to independent components (Oizumi et al., 2014).

Aim #2 took the form of the paper: "Causal emergence and the spatiotemporal scale of consciousness" by Hoel, Albantakis, Marshall, & Tononi (in prep) and is included as Chapter 2. In order to show that Φ^{Max} can occur at a macro spatiotemporal scale, it was necessary to fully characterize the cause-effect structure of a system and its irreducibility, and to do so at each possible spatiotemporal grain. Again we considered

similar small "neural" systems and measured, across all possible spatiotemporal scales, integrated information values for all subsets (rather than the *effective information* for the system as a whole). We also made use of the theoretic notions developed in Aim #1, such as how to apply a "macro causal intervention" to systems, the types of groupings allowable for use in generating coarse-grains, and also the decomposition of *cause-effect information* into the causal properties of state-space size, determinism, and degeneracy (Hoel et al., 2013). The Aim made use of the existent Python code from Aim #1 (which analyzes systems of logic gates across all possible spatiotemporal grains), which was eventually integrated into the Python code developed to measure Φ^{Max} (PyPhi).

The Aim gave a demonstration that Φ can reach a maximum at a macro spatial scale, at a macro temporal scale, and at a macro spatiotemporal scale. Building on this we identified how the causal structure of a system influences Φ at different levels and what causal properties are important for macro beating micro. It was shown that size of causal relationships, their selectivity (determinism and degeneracy), and the shift in their selectivity following a partition, together all contribute to integrated information. Continuing this focus, we were able to identify the kinds of network organizations which lead to a macro Φ^{Max}, such as highly degenerate or noisy systems.

Aim #2 accomplished extending Aim #1 by assessing Φ directly and thus showing that it is indeed possible that macro causal models can have greater capacity to integrated information than micro causal models. Furthermore the paper clarified the relationship between Φ and the causal structure and organization of a system.

Synaptic refinement, sleep-slow wave activity, and information integration

It is well established that, during infancy and adolescence, cortical connectivity is extensively reorganized through synaptogenesis, pruning, and synaptic plasticity. Brain development, and correspondingly, cognitive development, is an extended process of fine-tuning the neural circuitry (Feinberg, 1983). Such fine-tuning can be seen in the adult visual cortex, in which the connection probability between neurons is related to the similarity of their responsiveness to visual stimuli, indicating specificity in local synaptic connections (Ko et al., 2011). Important for responsiveness and information transmission, this type of organization allows for the sorting and grouping of sensory information along degrees of difference and similarity. However, this type of like-to-like local connectivity, which helps organize the cortex, does not exist at eye opening (Ko et al., 2013). Instead it develops over time, going through a critical period, beginning prior to adolescence, when a large amount of restructuring of connectivity is performed (synaptic refinement). This process that correlates with cognitive maturation (Buchmann et al., 2011).

Also during this period from pre- to post-adolescence there is thought to occur synaptic pruning: a decline in the total number of synapses and corresponding gray matter volume (Blakemore and Choudhury; Lenroot and Giedd 2006). At the same time there has been observed a mark decrease in sleep slow wave activity, which has been seen across adolescence in multiples species (Kurth et al., 2010; Nelson et al., 2013). It has been suggested that the decline in slow wave activity may be caused by synaptic pruning (Feinberg 1983; Campbell and Feinberg 2009) although direct evidence for pruning is minimal (Paus et al. 2008; Petanjek et al. 2011). An alternative hypothesis is that the drop in slow wave activity may be caused by synaptic refinement.

A convenient starting point to investigate the relationship between synaptic refinement and changes in slow wave activity is provided by a large-scale neural model of the primary visual cortex and associated thalamic regions (S. K. Esser, Hill, & Tononi, 2005, 2007, 2009; Hill & Tononi, 2005; Olcese, Esser, & Tononi, 2010). Simulations using this model have provided realistic approximations of evoked cortical activity during the presentation of visual stimuli, of spontaneous activity in wakefulness and sleep, of cortical responses to transcranial magnetic stimulation (TMS), and of changes in synaptic strength occurring with learning during wakefulness and sleep. For the work in this thesis the model has been adapted to investigate the effects of synaptic refinement in comparison to experimental data (Ko et al., 2013) by increasing the number of orientation selectivities and by implementing two different stages of development of its synaptic organization. These are a pre-refinement stage (child-like, or immature) in which connections are distributed at random in a homogenous manner across both topography and orientation selectivity; and a post-refinement (adult-like, or mature) stage in which connections to nearby topographic locations and similar orientation selectivities become more numerous, and other connections are pruned (Hoel, Albantakis, Cirelli, & Tononi, under review). Moreover, the model can be run in both a wake and a sleep mode, to assess whether synaptic refinement can account for changes of spontaneous activity, especially the observed decrease in sleep slow wave activity.

Synaptic refinement may be a brain-wide developmental principle: after the neonatal formation of connections which are (at the local level) connected randomly, or homogenously, these connections are then locally reorganized into specific functional patterns during activity-dependent learning, primarily during critical developmental

periods (Cramer, 1995). Thus, the model also provides an opportunity to relate brain organization with the development of consciousness: IIT proposes that certain network structures and motifs are associated with high levels of Φ, and that development should increase Φ in the cortex (Tononi, 2004) and a model of synaptic refinement could be used to test this hypothesis (see Future Directions).

Aim #3

The goal of this Aim was to implement a large-scale thalamocortical model before and after the establishment of selective local connectivity (synaptic refinement) and to show that synaptic refinement can account for the marked reduction in slow wave activity seen during brain development in multiple species. Aim #3 took the form of the paper: "Synaptic refinement during development and its effect on slow wave activity – a computational study" by Hoel, Albantakis, Cirelli, & Tononi (under review at *The Journal of Neurophysiology*). It is included as Chapter 3.

In order to assess how such circuits develop, and what effects follow such developmental structural changes, I optimized a previously existent model: a large-scale thalamocortical model of the primary visual cortex and associated thalamic regions, previous versions of which accurately represent cortical activity during wake and sleep (S. K. et al., 2005, 2007, 2009; Hill & Tononi, 2005; Olcese et al., 2010). Using the model, we performed simulations to investigate the effects of synaptic reorganization during development. This was done by reorganizing the model's connections from a pre-refined state (child-like, or immature) to a more developed, post-refined state (adult-like, or mature), in a manner consistent with physiological data (Ko et al., 2013).

Our simulations revealed several important effects of synaptic refinement on both the waking and sleeping activity of the thalamocortical model, primarily in the form of a reduction of slow wave activity (defined as amount of EEG delta power) during sleep. We also showed how synaptic refinement and its effects on slow wave activity can be produced by spike-timing dependent plasticity (STDP) during learning in wake (visual stimuli) and global synaptic renormalization in sleep. Moreover, our study can account for the experimental observation, in multiple species, of a marked decrease in slow wave activity over a similar period of development (Campbell and Feinberg, 2009).

References

Albantakis L, Hoel EP, Tononi G. The " neural code " from the intrinsic perspective : Quantifying causal power at different spatio-temporal scales. *Front. Comput. Neurosci. Conf. Abstr. Bernstein Conf. 2012*. .

Bateson G. *Steps to an Ecology of Mind*. University of Chicago Press., 1972.

Blakemore S-J, Choudhury S. Development of the adolescent brain: implications for executive function and social cognition. *J Child Psychol Psychiatry* 47: 296–312.

Borst A, Theunissen FE. Information theory and neural coding. *Nat Neurosci* 2: 947–57, 1999.

Buchmann A, Ringli M, Kurth S, Schaerer M, Geiger A, Jenni OG, Huber R. EEG sleep slow-wave activity as a mirror of cortical maturation. *Cereb Cortex* 21: 607–15, 2011.

Buxhoeveden D, Casanova M. The minicolumn hypothesis in neuroscience. *Brain* 125: 935–951, 2002.

Campbell IG, Feinberg I. Longitudinal trajectories of non-rapid eye movement delta and theta EEG as indicators of adolescent brain maturation. *Proc Natl Acad Sci U S A* 106: 5177–80, 2009.

Cramer K. Activity-dependent remodeling of connections in the mammalian visual system. *Curr Opin Neurobiol* 5: 106–111, 1995.

Esser SK, Hill S, Tononi G. Breakdown of effective connectivity during slow wave sleep: investigating the mechanism underlying a cortical gate using large-scale modeling. *J Neurophysiol* 102: 2096–111, 2009.

Esser SK, Hill SL, Tononi G. Modeling the effects of transcranial magnetic stimulation on cortical circuits. *J Neurophysiol* 94: 622–39, 2005.

Esser SK, Hill SL, Tononi G. Sleep homeostasis and cortical synchronization: I. Modeling the effects of synaptic strength on sleep slow waves. *Sleep* 30: 1617–30, 2007.

Feinberg I. Schizophrenia: caused by a fault in programmed synaptic elimination during adolescence? *J Psychiatr Res* 17: 319–34, 1983.

Georgopoulos A, Schwartz A, Kettner R. Neuronal population coding of movement direction. *Science (80-)* 233: 1416–1419, 1986.

Hill S, Tononi G. Modeling sleep and wakefulness in the thalamocortical system. *J Neurophysiol* 93: 1671–98, 2005.

Hoel EP, Albantakis L, Tononi G. Quantifying causal emergence shows that macro can beat micro. *Proc Natl Acad Sci U S A* 110: 19790–5, 2013.

Hoel EP, Hogan EP. The network properties of conscious experience: 'small worlds', clusters, and functional connectivity. *Association for the Scientific Study of Consciousness* 14, 2010.

Kandel ER, Markram H, Matthews PM, Yuste R, Koch C. Neuroscience thinks big (and collaboratively). *Nat Rev Neurosci* 14: 659–64, 2013.

Kim J. *Mind in a physical world: An essay on the mind-body problem and mental causation*. Cambridge, MA: MIT Press, 2000.

Ko H, Cossell L, Baragli C, Antolik J, Clopath C, Hofer SB, Mrsic-Flogel TD. The emergence of functional microcircuits in visual cortex. *Nature* 496: 96–100, 2013.

Ko H, Hofer SB, Pichler B, Buchanan K a, Sjöström PJ, Mrsic-Flogel TD. Functional specificity of local synaptic connections in neocortical networks. *Nature* 473: 87–91, 2011.

Kumar A, Rotter S, Aertsen A. Spiking activity propagation in neuronal networks: reconciling different perspectives on neural coding. *Nat Rev Neurosci* 11: 615–27, 2010.

Kurth S, Jenni OG, Riedner BA, Tononi G, Carskadon MA, Huber R. Characteristics of sleep slow waves in children and adolescents. *Sleep* 33: 475–80, 2010.

Lenroot RK, Giedd JN. Brain development in children and adolescents: insights from anatomical magnetic resonance imaging. *Neurosci Biobehav Rev* 30: 718–29, 2006.

London M, Roth A, Beeren L, Häusser M, Latham P. Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex. *Nature* 466: 123–127, 2010.

Markram H. The blue brain project. *Nat Rev Neurosci* 7: 153–160, 2006.

Marom S. Neural timescales or lack thereof. *Prog Neurobiol* 90: 16–28, 2010.

Mashour GA. Top-down mechanisms of anesthetic-induced unconsciousness. *Front Syst Neurosci* 8: 115, 2014.

Nelson AB, Faraguna U, Zoltan JT, Tononi G, Cirelli C. Sleep patterns and homeostatic mechanisms in adolescent mice. *Brain Sci* 3: 318–43, 2013.

Oizumi M, Albantakis L, Tononi G. From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Comput Biol* 10: e1003588, 2014.

Olcese U, Esser SK, Tononi G. Sleep and synaptic renormalization: a computational study. *J Neurophysiol* 104: 3476–93, 2010.

Panzeri S, Brunel N, Logothetis NK, Kayser C. Sensory neural codes using multiplexed temporal scales. *Trends Neurosci* 33: 111–20, 2010.

Paus T, Keshavan M, Giedd JN. Why do many psychiatric disorders emerge during adolescence? *Nat Rev Neurosci* 9: 947–57, 2008.

Petanjek Z, Judaš M, Šimic G, Rasin MR, Uylings HBM, Rakic P, Kostovic I. Extraordinary neoteny of synaptic spines in the human prefrontal cortex. *Proc Natl Acad Sci U S A* 108: 13281–6, 2011.

Quiroga RQ, Kreiman G, Koch C, Fried I. Sparse but not “grandmother-cell” coding in the medial temporal lobe. *Trends Cogn Sci* 12: 87–91, 2008.

Sporns O, Tononi G, Kötter R. The human connectome: A structural description of the human brain. *PLoS Comput Biol* 1: e42, 2005.

Tononi G, Sporns O. Measuring information integration. *BMC Neurosci* 4: 31, 2003.

Tononi G. An information integration theory of consciousness. *BMC Neurosci* 5: 42, 2004.

Tononi G. Consciousness as integrated information: a provisional manifesto. *Biol Bull* 215: 216–242, 2008.

Tononi G. Integrated Information Theory of Consciousness: An Updated Account. *Arch Ital Biol* 150: 56–90, 2012a.

Tononi G. *Phi: A Voyage from the Brain to the Soul*. Knopf Doubleday Publishing Group, 2012b.

Vijayan S, Hale GJ, Moore CI, Brown EN, Wilson M. Activity in the barrel cortex during active behavior and sleep. *J Neurophysiol* 103: 2074–84, 2010.

CHAPTER ONE

Quantifying causal emergence shows that macro can beat micro

Erik P Hoel[1], Larissa Albantakis[1], Giulio Tononi[1]
[1] Department of Psychiatry, University of Wisconsin, Madison, WI, USA

Published in:
The Proceedings of the National Academy of Sciences 2013; 110(49): 19790-19795

Abstract

Causal interactions within complex systems can be analyzed at multiple spatial and temporal scales. For example, the brain can be analyzed at the level of neurons, neuronal groups, and areas, over tens, hundreds, or thousands of milliseconds. It is widely assumed that, once a micro level is fixed, macro levels are fixed too, a relation called supervenience. It is also assumed that, while macro descriptions may be convenient, only the micro level is causally complete, since it includes every detail, thus leaving no room for causation at the macro level. However, this assumption can only be evaluated under a proper measure of causation. Here, we employ a measure (effective information, EI, (Tononi & Sporns, 2003)) that depends on both the effectiveness of a system's mechanisms and the size of its state space: *EI* is higher the more the mechanisms constrain the system's possible past and future states. By measuring *EI* at micro and macro levels in simple systems whose micro mechanisms are fixed, we show that for certain causal architectures *EI* can peak at a macro level in space and/or time. This happens when coarse-grained macro mechanisms are more effective (more deterministic and/or less degenerate) than the underlying micro mechanisms, to an extent that overcomes the smaller state space. Thus, although the macro level supervenes upon the micro, it can supersede it causally, leading to genuine causal emergence - the gain in *EI* when moving from a micro to a macro level of analysis.

Significance

Properly characterizing emergence requires a causal approach. Here we construct causal models of simple systems at micro and macro spatiotemporal scales and measure their

causal effectiveness using a general measure of causation (*effective information*). *EI* is dependent on the size of the system' state-space and reflects key properties of causation (selectivity, determinism, and degeneracy). While in the example systems, the macro mechanisms are completely specified by their underlying micro mechanisms, *EI* can nevertheless peak at a macro spatiotemporal scale. This approach leads to a straightforward way of quantifying causal emergence as the supersedence of a macro causal model over a micro one.

Introduction

In science it is usually assumed that, the better one can characterize the detailed causal mechanisms of a complex system, the more one can understand how the system works. At times it may be convenient to resort to a macro level description, either because not all the micro level data are available, or because a rough model may suffice for one's purposes. However, a complete understanding of how a system functions, and the ability to predict its behavior precisely, would seem to require the full knowledge of causal interactions at the micro level. For example, the brain can be characterized at a macro scale of brain regions and pathways, a meso scale of local populations of neurons such as minicolumns and their connectivity, and a micro scale of neurons and their synapses (Sporns et al., 2005). With the goal of a complete mechanistic understanding of the brain, ambitious programs have been launched with the aim of modeling its micro scale (Markram, 2006).

The reductionist approach common in science has been successful not only in practice, but has also been supported by strong theoretical arguments. The chief argument

starts from the intuitive notion that, when the properties of *micro* level physical mechanisms of a system are fixed, so are the properties of all its *macro* levels – a relation called *supervenience* (Davidson, 1980). In turn, this relation is usually taken to imply that the micro mechanisms do all the causal work, i.e. the micro level is causally complete. This leaves no room for any causal contribution at the macro level, otherwise there would be *multiple causation* (Kim, 1993). This *causal exclusion* argument is often applied to argue against the possibility for mental causation above and beyond physical causation (Kim, 2000), but it can be extended to all cases of supervenience, including the hierarchy of the sciences (Bontly, 2002).

Some have nevertheless argued for the possibility that genuine emergence can occur. Purported examples go all the way from the behavior of flocks of organisms (Seth, 2008) to that of ant colonies (Hölldobler and Wilson, 2009), brains (Sperry, 1983), and of human societies (Sawyer, 2005). Unfortunately, it remains unclear what would qualify some systems as truly emergent and others as reducible to their micro elements. Also, most arguments in favor of emergence have been qualitative (Broad and Paul, 1925). A convincing case for emergence must demonstrate that higher levels can be causal above and beyond lower levels (*causal emergence*). So far, the few attempts to characterize emergence quantitatively (Bar-Yam, 2004) have not been based on causal models.

Here we make use of simple simulated systems, including neural-like ones, to show quantitatively that the macro level can causally supersede the micro level, i.e. causal emergence can occur. We do so by perturbing each system through its entire repertoire of possible causal states (*counterfactuals*, in the general sense of alternative possibilities) and evaluating the resulting effects using *effective information* (*EI*) (Tononi

and Sporns, 2003). *EI* is a general measure for causal interactions because it uses perturbations to capture the effectiveness/selectivity of the mechanisms of a system in relation to the size of its state space. As will be pointed out, *EI* is maximal for systems that are deterministic and not degenerate, and decreases with noise (causal divergence) and/or degeneracy (causal convergence).

For each system we completely characterize the causal mechanisms at the micro level, fixing what can happen at any macro level (supervenience). Macro levels are defined by coarse-graining the micro elements in space and/or time, and this mapping defines the repertoire of possible causes and effects at each level. By comparing *EI* at different levels we show that, depending on how a system is organized, causal interactions can peak at a macro rather than at a micro spatiotemporal scale. Thus, the macro may be causally superior to the micro even though it supervenes upon it. Evaluating the changes in *EI* that arise from coarse or fine graining a system provides a straightforward way of quantifying both emergence and reduction.

Theory

In what follows, we consider discrete systems *S* of connected binary micro elements that implement logical functions (mechanisms) over their inputs. We will first introduce a state-dependent measure of causation, the *cause* and *effect information* of a single system state s_0, before we describe the state-independent *effective information (EI)* of the system *S*.

State-dependent causal analysis. The micro mechanisms of *S* specify its state-to-state transition probability matrix (TPM) at a micro time step *t*. Building upon the perturbational framework of causal analysis developed by Judea Pearl (Pearl, 2000), the TPM can be obtained by perturbing *S* at t_0 (Tononi and Sporns, 2003) into all possible *n* initial states with equal probability 1/*n* ($do(S = s_i)$ " i Î 1□ n). Perturbing the system in this way corresponds to the unconstrained repertoire (probability distribution) of possible causes U^C, and determines the probability of the resulting states at t_{+1}, corresponding to the unconstrained repertoire of possible effects U^E. While U^C is thus identical to the uniform distribution *U* (with $p(s) = 1/n$, " s Î S), U^E is typically not uniform. A current system state $S=s_0$ is associated with the probability distribution of past states that could have caused it (*cause repertoire* $S_P|s_0$, obtained by Bayes rule), and the probability distribution of future states that could be its effects (*effect repertoire* $S_F|s_0$). A system's mechanisms and current state thus constrain both the repertoire of possible causes U^C and that of possible effects U^E. An informational measure of the causal interactions in the system (Albantakis et al., 2013) can then be defined as the difference (here Kullback-Leibler Divergence, D_{KL} (Kullback, 1997)) between the constrained and unconstrained distributions:

$$\text{Cause information}(s_0) = D_{KL}\left(\left(S_P|s_0\right), U^C\right), \quad \text{Effect information}(s_0) = D_{KL}\left(\left(S_F|s_0\right), U^E\right).$$

Cause/effect information depends on two properties: 1.) the *size* of the system's state space (repertoire of alternatives), since both are bounded by $log_2(n)$; 2.) the *effectiveness* of the system's mechanisms in specifying past and future states. To isolate effectiveness from size, we define the following normalized coefficients:

$$\text{Cause coefficient}(s_0) = \frac{\text{Cause Information}(s_0)}{\log_2(n)}, \quad \text{Effect coefficient}(s_0) = \frac{\text{Effect Information}(s_0)}{\log_2(n)}.$$

The cause coefficient describes to what extent a state is sufficient to specify its past causes, and the effect coefficient indicates how necessary a state is to specify its future effects (see Fig. 1.1B). In turn, the effect coefficient itself is a function of two terms, determinism and degeneracy (see Appendix I S1 for derivation):

$$\begin{aligned}\text{Effect coefficient}(s_0) &= \text{Determinism coefficient}(s_0) - \text{Degeneracy coefficient}(s_0) \\ &= \frac{1}{\log_2(n)} \sum_{s_F \in U^E} p(s_F|s_0)\log_2\left(n \times p(s_F|s_0)\right) - \frac{1}{\log_2(n)} \sum_{s_F \in U^E} p(s_F|s_0)\log_2\left(n \times p(s_F)\right)\end{aligned}$$

.

The *determinism* coefficient is the difference $D_{KL}\left(\left(S_F|s_0\right),U\right)$ between the effect repertoire and the uniform distribution (*U*) of system states, divided by $log_2(n)$, and measures how deterministically (reliably) s_0 leads to the future state of the system: it is '1' (complete determinism) when the current state leads to a single future state with probability p=1, and is '0' (complete indeterminism or noise) if it could be followed by every future state with $p=1/n$. The *degeneracy* coefficient measures to what degree there is deterministic convergence (not due to noise) from other states onto the future states specified by s_0. In broad terms, degeneracy refers to multiple ways of deterministically achieving the same effect or function (Edelman, 1987; Tononi et al., 1999). The degeneracy coefficient is '1' (complete degeneracy) when s_0 specifies the same future state as all other states, and '0' when s_0 specifies a unique future state (no degeneracy).

Both cause and effect coefficients are minimal (0) in a completely noisy or completely degenerate system (Fig. 1.1C,D) and maximal (1) in a deterministic, non-degenerate system (see Appendix I S3). The contribution of a single state to the system's

determinism and degeneracy are best demonstrated by decomposing the effect coefficient. While the cause coefficient also reflects the degeneracy and determinism of the system, it is not subdivided further here.

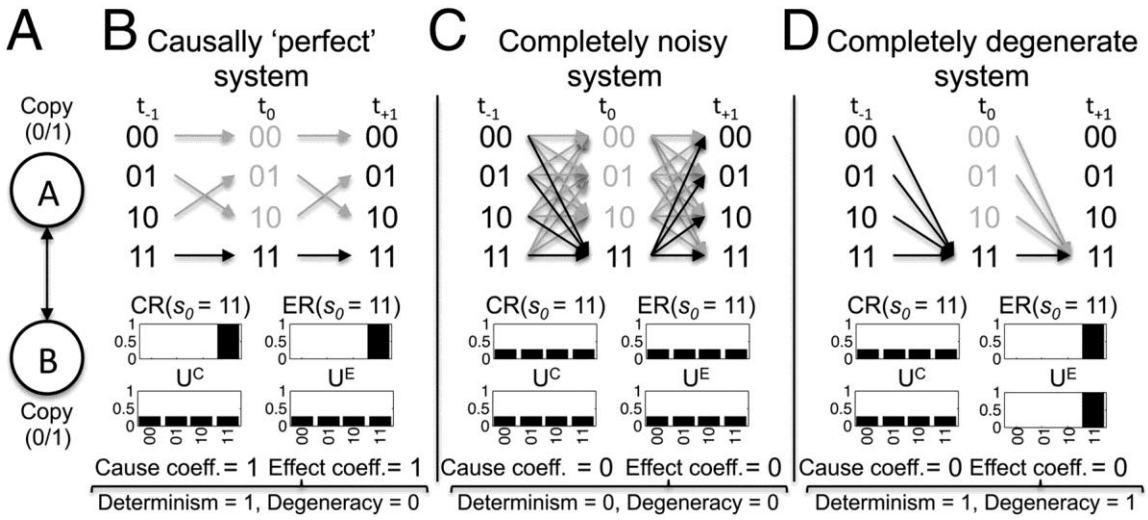


Figure 1.1. Cause and effect coefficients in example systems with different causal architectures. (A) The systems consist of 2 interconnected binary COPY gates with possible states '0' and '1'. (B) A causally perfect system, in which each state has one cause and one effect. Thus, s_0=[11] has a cause and effect coeff. of 1. Moreover, there is no divergence (determinism coeff=1) and no convergence (degeneracy coeff.=0). (C,D) In both the completely indeterministic and completely degenerate systems, state s_0=[11] is completely insufficient to specify past system states and completely unnecessary to specify future states (cause and effect coeff.=0). Note that the degeneracy coefficient is 0 in the completely noisy system, since all convergence is due to noise alone.

State-independent causal analysis. A state-independent informational measure of a system's causal architecture can be obtained by taking the expected value of cause or effect information over all system states, a quantity called *effective information* (*EI*):

$$EI(S) = \langle \text{Cause Information}(s_0) \rangle = \sum_{s_0 \in U^E} p(s_0) D_{KL}\left(\left(S_P | s_0\right), U^C\right) = \langle \text{Effect Information}(s_0) \rangle = \frac{1}{n} \sum_{s_0 \in U^C} D_{KL}\left(\left(S_F | s_0\right), U^E\right)$$

The two terms are identical, since the system is assumed to be time invariant ($\langle t_{-1} \to t_0 \rangle = \langle t_0 \to t_{+1} \rangle$), and cause and effect information are related via Bayes' rule (see Text S2 for details). *EI* is also the mutual information (*MI*) between all possible causes and their effects, $MI(U^C;U^E)$ (Text S2).

As a measure of causation, *EI* captures how effectively (deterministically and uniquely) causes produce effects in the system, and how selectively causes can be identified from effects. As with the state-dependent measures, the effectiveness (*Eff*) of the causal interactions within a system can be captured by normalizing *EI* by the system's size: $Eff(S) = EI(S) / \log_2(n)$. Also as in the state-independent case, effectiveness can be split into two components, determinism and degeneracy:

$$\begin{aligned} Eff(S) &= \langle \text{Determinism coefficient}(s_0) \rangle - \langle \text{Degeneracy coefficient}(s_0) \rangle \\ &= \left\langle D_{KL}\left(\left(S_F | s_0 \right), U \right) \right\rangle / \log_2(n) - \quad D_{KL}\left(U^E | U \right) / \log_2(n) \end{aligned} .$$

Thus, *Eff(S)*=1 if *EI* is maximal for a given system size, and decreases with indeterminism (divergence due to noise) or degeneracy (deterministic convergence), with *Eff(S)*=0 for completely noisy or degenerate systems (Fig. 1.1C,D). In a system with perfect effectiveness (Fig. 1.1B), each cause has a unique effect, and each effect has a unique cause. Thus, such a system (where *Eff*(*S*)=1) is perfectly retrodictive/predictive, in the sense that not only the unique future trajectory, but also the unique past trajectory of all states can be deduced from the TPM (complete causal reversibility).

Levels of analysis. A finite, discrete system *S* can be considered at various levels, from the most fine-grained micro causal model S_m through various coarse grained causal models S_M. All macro levels S_M are assumed to be *supervenient* on the micro level S_m:

given the micro elements of S_m and the causal relationships between them, all other members of {**S**} - the set of all possible causal models of system *S* - are fixed as well (Stalnaker, 1996). Although S_m fixes S_M, any S_M may be fixed by a number of different lower level descriptions, a property known as *multiple realizability* (Fodor, 1974).

Groupings. Micro elements are binary and labeled by Latin letters {A,B,C…}, macro elements by Greek letters {α,β,γ…}. Micro states are labeled {1,0} and macro states {'on', 'bursting', 'quiet'…). Micro elements can be grouped into macro elements spatially, temporally, or both. Micro states are grouped into macro states through a mapping $M : S_m \rightarrow S_M$. The mapping must be exhaustive and disjunctive over micro elements (all the states of one micro element must be mapped to the states of the same macro element; note that a macro element can consist of a single micro element as long as the state space of the system is reduced). Moreover, the mapping must be such that no micro level information is available at the macro level, (the identity of the micro elements grouped into a macro element is lost). For example, the grouping of the 4 states of 2 micro elements into the 2 states of 1 macro element as [00, 01, 10]=off, [11]=on] is permitted, whereas the grouping [[00, 01], [10, 11]] is not, since distinguishing 01 from 10 requires knowing the identity of the micro elements.

Level-specific perturbations. Causal analysis at the micro level S_m, requires setting *S* into all possible micro states with equal probability (i.e. testing all micro alternatives) and determining the resulting effects. When moving to a macro level S_M, *S* must similarly be set into all possible macro states with equal probability (i.e. testing all macro

alternatives). To causally assess any macro state, then, one must set S into all the n_{micro} micro states $\{s_m\}$ that are grouped into the corresponding macro state s_M, and average over the effects. This is done using a *macro perturbation*:

$do(S_M = s_M) = \frac{1}{n_{micro}} \sum_{s_{m,i} \in s_M} do(S_m = s_{m,i})$. Using such macro perturbations one can obtain cause/effect information and *EI* for every coarse grain of S_m. *EI* at each macro level is then equivalent to the *MI* between the set of macro causes and their macro effects.

Causal emergence/reduction. Finally, by assessing *EI(S)* over all coarse grains of S_m, one can ask at which level of $\{\mathbf{S}\}$ causation reaches a maximum. This provides an analytical definition of *causal emergence (CE)*, expressed in bits: $CE = EI(S_M) - EI(S_m)$. Thus, if *EI(S)* is maximal for a macro level S_M rather than the micro level S_m, then $CE>0$ and causal emergence occurs. If for every macro level $CE<0$, causal reduction holds. While the focus here is on emergence/reduction relative to the micro level S_m, the above measure can of course be used to compare different macro levels.

As also mentioned above, *EI(S)* depends on both the size of the system's repertoire of states and on the effectiveness of its mechanisms. When moving from one system level to another, both terms change as the state-space becomes smaller or larger, and the individual states become more or less selective with respect to the past, and more or less determined or degenerate with respect to the future. The respective informational contributions of repertoire size and effectiveness to $\Delta EI(S)$ can be expressed separately as:

$\Delta I_{Eff} = \left(Eff(S_M) - Eff(S_m)\right) \cdot \log_2(n_M)$, $\Delta I_{Size} = Eff(S_m) \cdot \left(\log_2(n_M) - \log_2(n_m)\right)$, where $n_{m/M}$ is the state repertoire size of $S_{m/M}$. It follows that $\Delta EI = \Delta I_{Eff} + \Delta I_{Size} = CE$. A positive ΔI_{Eff} can thus be due to the macro reducing the degeneracy of the micro level, increasing the determinism of the micro level, or both. Notably, coarse graining the micro level S_m into macro level S_M implies that ΔI_{Size} is always negative. Hence, for causal emergence to occur ($EI(S_M) > EI(S_m)$), the increase in effectiveness ΔI_{Eff} must outweigh the decrease in ΔI_{Size}.

Results. Causal analysis was performed across all coarse grains of a system (only the S_M with maximal *EI(S)* is shown in the figures) with a custom-made Python program. Data plots were created using MATLAB. Below, we consider examples of spatial, temporal, and spatiotemporal emergence (see Appendix I Fig. AI.1 for an example of spatial reduction).

Spatial causal emergence. As a proof-of-principle example, consider a system of 4 binary elements S_m={ABCD} (Fig. 1.2A). Each micro mechanism is an AND-gate (2 inputs) operating over some intrinsic noise. The 16x16 S_m TPM was constructed by setting the system into all possible micro states from [0000] to [1111] with equal probability (Fig. 1.2B). At the micro level S_m, effective information *EI(S)*=1.15 bits, out of maximally 4 bits, with effectiveness $Eff(S_m)$=0.29. The macro level S_M (Fig. 1.2D), composed of 2 elements {α, β}, each with states {'on', 'off'}, is a coarse graining of S_m as defined by the mapping ***M*** in Fig. 1.2C. The 4x4 S_M TPM was obtained by setting the system into all possible macro states from [off, off] to [on, on] with equal probability

(Fig. 1.2E). For the macro level, $EI(S_M)$=1.55 bits, higher than $EI(S_m)$=1.15 bits. Thus, $CE(S)$=0.40 bits, demonstrating that in this case the macro S_M beats the micro S_m and constitutes the optimal causal model of system S. This is because the TPM for S_M is much closer to perfect effectiveness ($Eff(S_M)$=0.78) and the increase in effectiveness gained by grouping ΔI_{Eff}=0.97 bits outweighs the loss in size ΔI_{Size}=-0.57 bits). In this example, the gain in effectiveness ΔI_{Eff} at the macro level comes primarily (91%) from counteracting noise (determinism coeff. (S_m)=0.34; (S_M)=0.78) and less so (9%) from reducing degeneracy (degeneracy coeff. (S_m)=0.05; (S_M)=0.006).

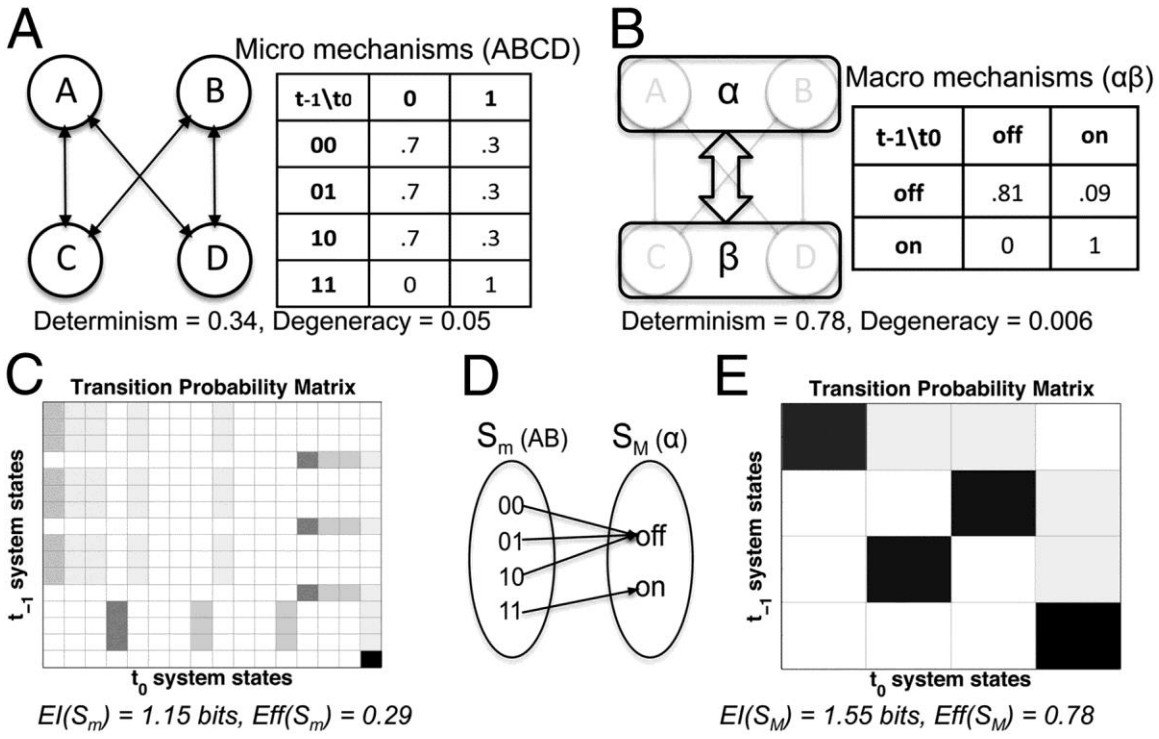


Figure 1.2. Spatial causal emergence (counteracting indeterminism). (A) The micro level S_m of system S is composed of identical noisy micro mechanisms. (B) The micro TPM. (C) A macro causal level S_M and its TPM are defined by the mapping $\boldsymbol{M}$ (shown for AB to α, CD to β is symmetric). (D) S_M and its macro mechanisms. (E) By reducing indeterminism and increasing effectiveness *Eff*, the macro beats the micro in terms of effective information *EI* despite the reduced repertoire size (***CE*=0.40 bits).**

The higher effectiveness of the macro level is also evident comparing S_m and S_M in a state-dependent manner. As an example, the cause/effect distributions for S_m in state {ABCD}=[0001] are compared to the corresponding S_M state {αβ}=[off, off] in Fig. 1.3. Comparing the cause/effect distributions of S_m=[0001] against the unconstrained repertoires (using D_{KL}) yields 0.83 bits of cause information and 0.43 bits of effect information. For the macro S_M, cause information is 2 bits and effect information 1.35 bits. The macro beats the micro because {αβ}=[off, off] is both more selective and more reliable than {ABCD}=[0001].

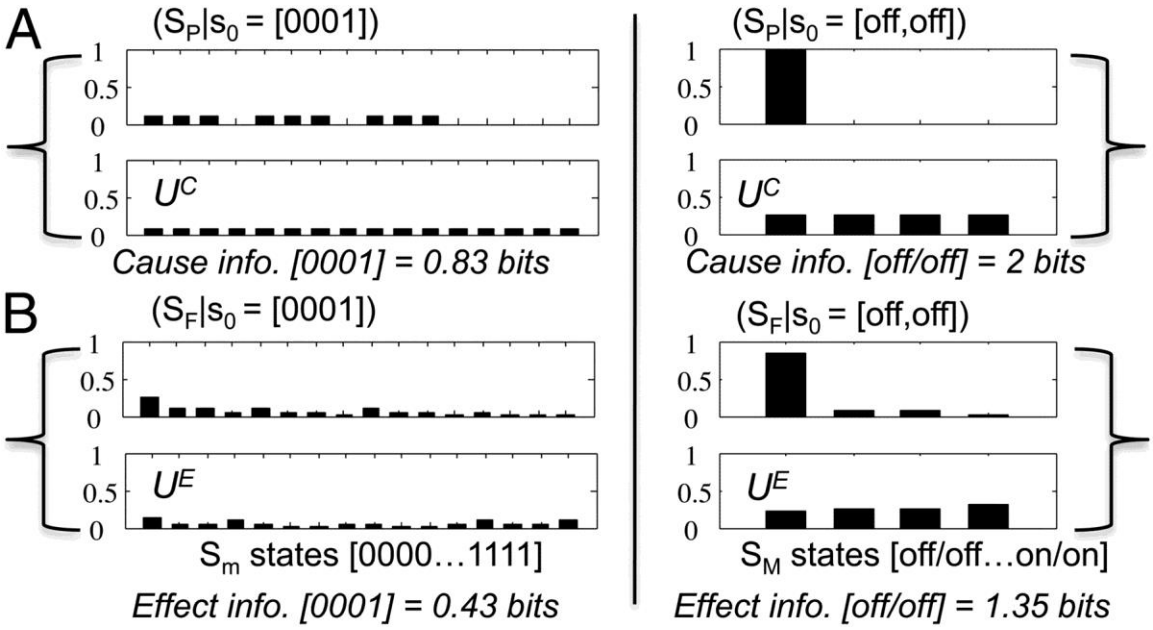


Figure 1.3. State dependent cause/effect information. (A) The cause information of S_m in micro state {ABCD}=[0001] is calculated as the difference (D_{KL}) between the cause repertoire of state [0001] and the unconstrained micro repertoire U^C (left). The cause information of S_M in the supervening macro state {αβ}=[off/off] (right) is the difference (D_{KL}), between the cause repertoire of [off/off] and the unconstrained macro repertoire U^C. (B) Effect information. The higher cause and effect information at the macro level is due to an increase in determinism and decrease in degeneracy, reflecting higher selectivity.

Causal emergence may arise not only from macro gains in determinism (as above), but also from reducing degeneracy. In Fig. 1.4, micro elements are deterministic AND gates connected in a way that ensures high degeneracy (A-F; Fig. 1.4A, determinism coeff.=1; degeneracy coeff.=0.6), resulting in $Eff(S_m)$=0.4 and $EI(S_m)$=2.43 bits (Fig. 1.4C). The optimal macro groups the 6 micro AND gates into 3 macro COPY gates (αβγ) (Fig. 1.4B). Both macro and micro are deterministic, but by eliminating degeneracy ΔI_{Eff}=1.79 bits > $-\Delta I_{Size}$=1.22 bits. As a result, $Eff(S_M)$=1, $EI(S_M)$=3 bits, and the macro emerges over the micro (*CE*=0.57 bits).

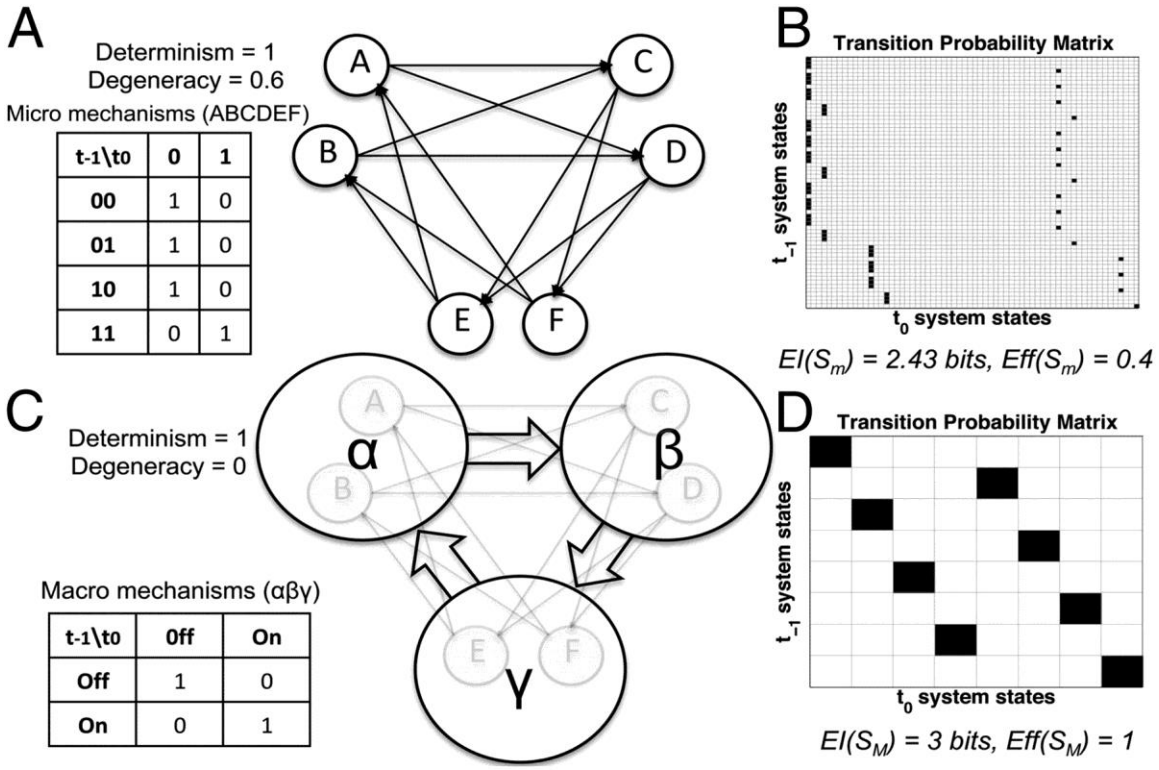


Figure 1.4. Spatial causal emergence (counteracting degeneracy). (A) A degenerate S_m with deterministic AND gates. (B) The cycle of AND gates is mapped onto a cycle of COPY gates at the macro level. (C) The deterministic but degenerate micro TPM. (D) The deterministic macro TPM with zero degeneracy. By eliminating degeneracy and achieving perfect effectiveness, the macro beats the micro (***CE*=0.57 bits**).

Temporal causal emergence

The same principles allowing for emergence through spatial groupings hold for temporal groupings, which coarse grain micro time steps (t_x) into macro time steps (T_x). The example in Fig. 1.5 shows micro elements that, upon receiving an input 'burst' of 2 spikes, respond with an output burst of 2 spikes. Thus, elements implement 2nd-order Markov mechanisms over both inputs and outputs (Fig. 1.5A). Fig. 1.5B shows that causal interactions assessed over 1 micro time step are weak ($EI(S_m)$=0.16 bits; $Eff(S_m)$=0.03) because they fail to capture the 2nd-order mechanisms. By contrast, causal analysis over 2 micro time steps (Fig. 1.5C) gives EI=1.38 bits and $Eff(S_m)$=0.34. The temporal grouping of micro into macro states $\alpha=\{A_t,A_{t+1}\}$ and $\beta=\{B_t,B_{t+1}\}$ (Fig. 1.5D) is analogous to the spatial grouping in Fig. 1.2: {00,01,10}={'off'} and {11}={'on'}. Over macro time steps, the system becomes fully deterministic and non-degenerate, $EI(S_M)$=2 bits, $Eff(S_M)$=1, and CE(S)=0.62 bits (Fig. 1.5D-E).

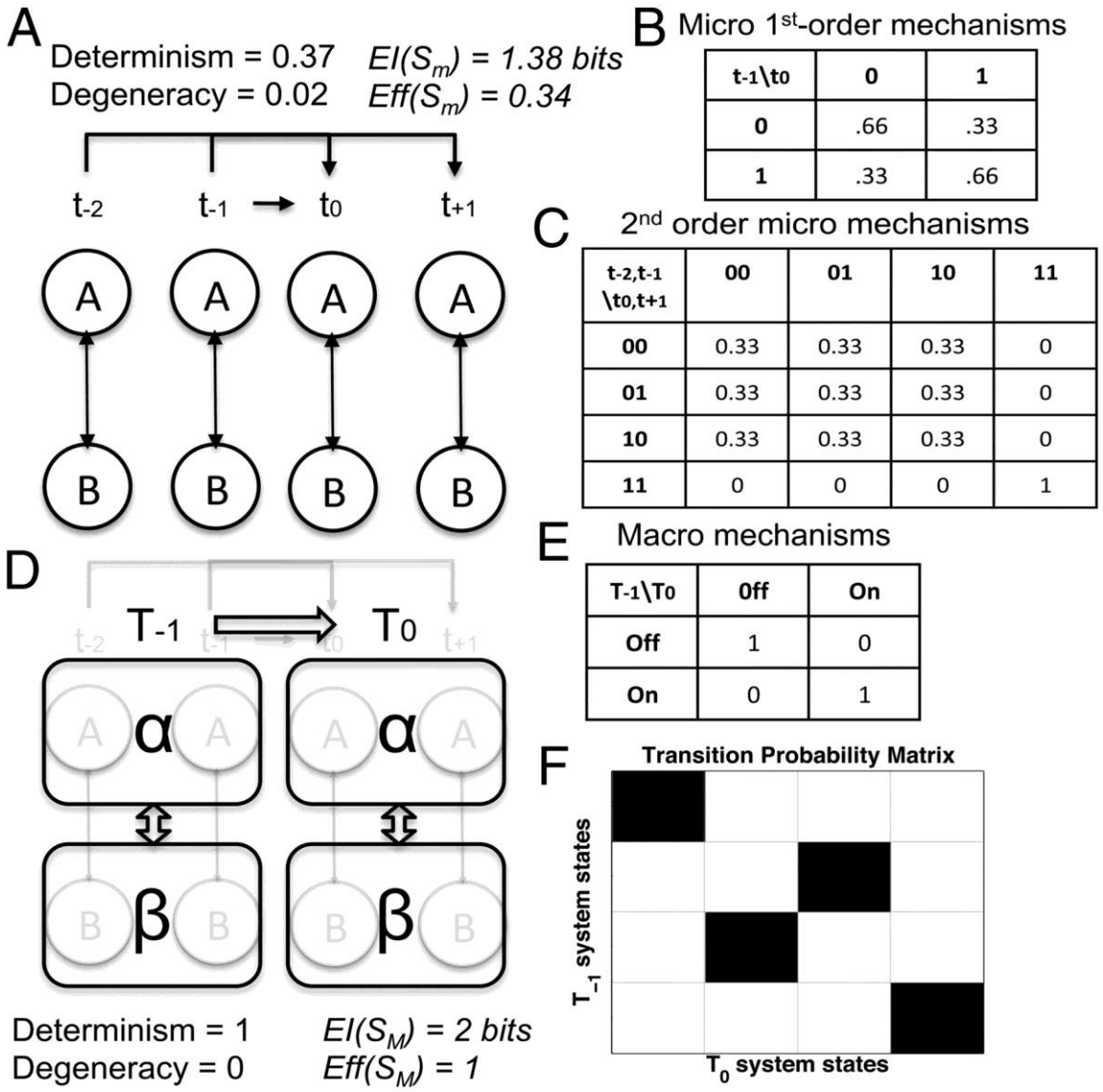


Figure 1.5: Temporal causal emergence. (A) S_m is composed of 2nd-order Markov mechanisms A and B: at t_0 each mechanism responds based on the inputs at t_{-2} and t_{-1}, and outputs over t_0 and t_{+1}. (B) Causal analysis over 1 micro time step gives an incomplete view of the system. (C) A causal analysis over 2 micro time steps reveals the 2nd-order Markov mechanisms. (D) The optimal macro system S_M groups 2 micro time steps into 1 macro time step for macro elements {α,β}. (D) Each coarse grained macro mechanism effectively corresponds to a deterministic COPY gate. (E) The macro 1-time step TPM S_M has $Eff(S_M)$=1, the micro 2-time step TPM has $Eff(S_m)$=0.34, ***CE*=0.62 bits**.

Spatiotemporal causal emergence

In general, emergence may occur simultaneously over space and time (Fig. 1.6). As in Fig. 1.5, the 9 neural-like micro elements in Fig. 1.6A are 2nd-order Markov mechanisms, integrating inputs and outputs over 2 micro time steps, $t_{-2}\,t_{-1}$, and $t_0\,t_{+1}$, respectively (cf. the longer time constants of NMDA receptors (Jahr and Stevens, 1990)). Moreover, in the examples above, the micro elements within a macro element were not connected and were causally equivalent. To demonstrate that this is not a requisite for causal emergence, in Fig. 1.6, the micro elements are fully connected and causally heterogeneous (self-connections not drawn). All elements are spontaneously active (1) with heterogeneous probabilities: p(A/D/G)=0.45; p(B/E/H)=0.5; p(C/F/I)=0.55. The elements are structured into 3 groups {ABC, DEF, GHI} due to different intra- and inter-group mechanisms: within each group, if the sum of intra-group connections Σ(intra)=0 (for 2 time steps), all elements stay 0 (for the next 2 time steps). However, if the sum of inter-group connections Σ(inter)=6 from one or both of the other 2 groups over 2 time steps (burst of synchronous activity), p(1) is raised by 0.5 for the next 2 time steps (see Fig. AI.2 for macro and micro TPMs of a spatial system with equivalent rules). At the macro level S_M (Fig. 1.6B), the 3 groups of neurons become macro elements, and 2 micro time steps (t_x) are grouped into 1 macro time step (T_x). In neural terms, these macro elements could represent ‘minicolumns’ having 3 states: 'inhibited' (all minicolumn neurons silent at T_x), 'receptive' (some firing at T_x), or 'bursting' (all firing at T_x). Macro causal interactions can be summarized as follows: if a macro element is 'inhibited', only receiving a 'burst' can move it to the 'receptive' or (more unlikely) the 'bursting' state, otherwise it stays 'inhibited'. As in previous examples, the coarse grained S_M has higher $EI(S_M)$=3.51 bits and $Eff(S_M)$=0.74 than S_m ($EI(S_m)$=0.59 bits; $Eff(S_m)$=0.033). In this case, spatiotemporal

causal emergence (*CE*(S)=2.92 bits) is due to an increase in determinism that far outweighs a slight increase in degeneracy and the decrease in size.

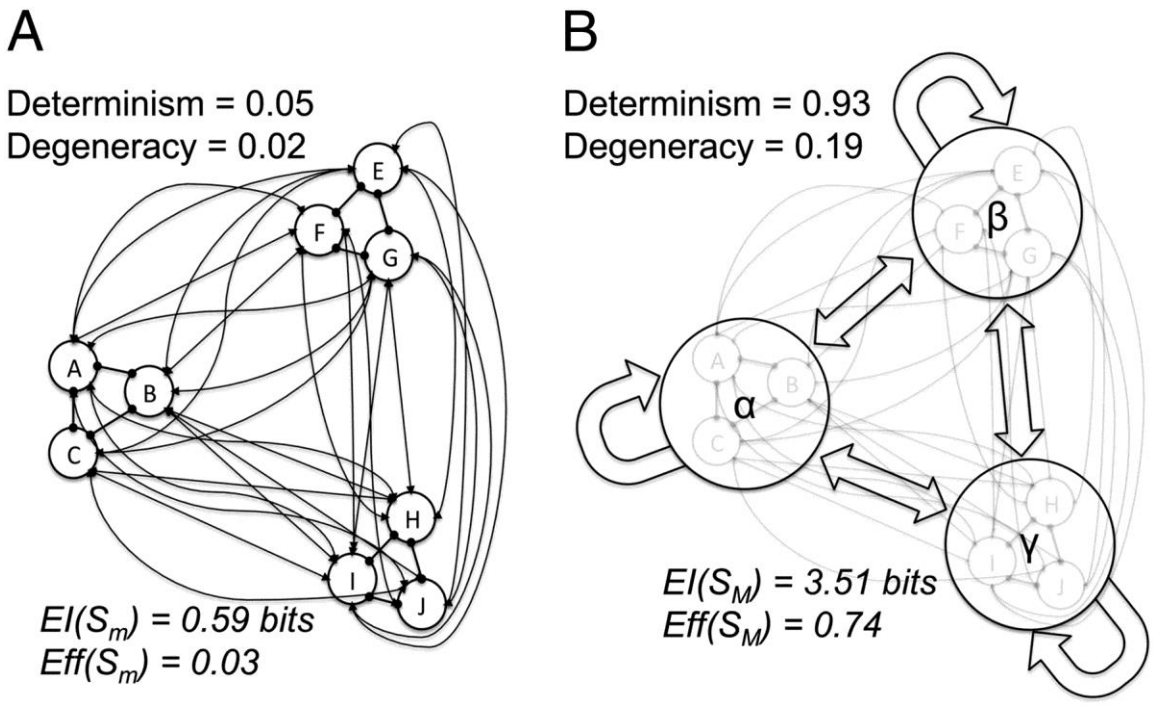


Figure 1.6. Spatiotemporal causal emergence. (A) A 'neuronal' system merging the temporal characteristics of the system in Fig. 1.5 with a differentiated spatial structure (Fig. AI.2). Regular and rounded arrows indicate inter and intra-group connections, respectively. (B) Each macro element receives inputs from itself and the other macro element. The macro level beats the micro level, leading to spatiotemporal emergence (***CE*(S)=2.92 bits**).

Discussion

This paper provides a principled way of assessing at which spatiotemporal grain size the causal interactions within a system reach a maximum. Causal interactions are evaluated by *effective information (EI)*, a measure that is sensitive both to the effectiveness of the system's mechanisms and to the size of its state space. Examples with simulated systems demonstrate that, after coarse-graining the micro mechanisms in both space and time, *EI*

can be higher at a macro level than at a micro level. In these cases, the macro mechanisms, rather than the micro ones, can be said to be doing the causal work within a system.

Effective information, effectiveness, and emergence. As shown here, *EI* corresponds to the *effectiveness* of a system's mechanisms multiplied by repertoire size, expressed in bits. Effectiveness *Eff(S)* is the average of the *effect coefficients* over all system states. The effect coefficient measures to what extent the current system state is necessary to specify the system's future state. This, in turn, is a function of *determinism* minus *degeneracy*. On the cause side, the equivalent to the effect coefficient is the *cause coefficient*, which measures to what extent the current state is sufficient to specify the system's past state. For a particular current state, cause and effect coefficients may differ: for example, a state may have many causes but only one effect. However the average of the effect coefficients over system states, i.e. effectiveness, corresponds to the average of the cause coefficients (weighted by the probability of the effects). In other words, within a time-invariant system the average selectivity of the causes corresponds to the average selectivity of the effects. Note that, in principle, other measures of causation that, like *EI*, reflect causal structure (selectivity, determinism, degeneracy) and system size, should demonstrate causal emergence as well.

The main result obtained in the simulations is that coarse-graining, both in space and in time, can yield a higher value of *EI*. This happens even though the micro has, by definition, a larger state space than the macro - an advantage with respect to *EI*. Given this inherent advantage of the micro, it is understandable why the default scientific

strategy for analyzing systems has been one of reduction. However, the examples presented above show that the inherent loss in *EI* due to the macro's smaller repertoire size can be offset if the macro achieves a greater gain in effectiveness. In turn, greater effectiveness stems from macro mechanisms constructed from their constituting micro mechanisms in such a way that, at the macro level, determinism is increased and/or degeneracy is decreased. Genuine causal emergence (*CE*) can then be said to occur whenever there is a gain in *EI* (*CE*>0) at the optimal macro level. If instead there is a loss in *EI* (*CE*<0), causal reduction is appropriate, and the micro level is the optimal level of causal analysis. The causal approach pursued here suggests that qualitative or non-causal accounts of emergence may have been hindered by not being able to characterize how and why a macro level can actually have greater causal effectiveness than a micro level (Bedau, 1997; Chalmers, 2006).

Micro macro mappings and repertoires of alternatives. The present approach makes it possible to compare causation at the micro and macro levels in a fair manner. First, the simulated examples are such that the macro supervenes strictly upon the micro: once the micro is defined, all macro levels are fixed. Specifically, no extra causal ingredients are added at the macro level, such as rules that apply to the macro only (Butterfield, 2012). Furthermore, the mapping of micro into macro elements is such that the identity of micro elements is lost, otherwise the macro level would have access to micro level information that could offset its reduced repertoire size. Finally, when causation is evaluated a uniform distribution of alternatives is imposed independently at the micro and macro levels. That is, all micro alternatives are equally likely at the micro level, but for macro

alternatives to be equally likely as well, the probability of the underlying micro perturbations must be modified by averaging the micro states that map into the same macro state. The modified distribution of micro perturbations yielding a uniform distribution of macro perturbations makes *EI* sensitive to the causal structure at each level, ultimately allowing the supervening macro *EI* to exceed the micro *EI*.

Emergence as an intrinsic property of a system. *EI* is a causal measure, since it requires perturbing the system in all possible ways and evaluating the resulting effects on the system. It is also an informational measure, since its value depends on the size of the repertoire of alternatives. Indeed, in the present approach causation and information are necessarily linked (Tononi, 2012), hence the term *effective information*. Finally, measuring *EI* reveals an *intrinsic* property of the system, namely the average effectiveness/selectivity of all possible system states with respect to the system itself. Effectiveness/selectivity can be assessed at multiple spatiotemporal grains, and the particular spatiotemporal grain at which *EI* reaches a maximum is again an intrinsic property of the system. This in no way precludes an observer from profitably investigating the system's properties at other macro levels, at the micro level, or at multiple levels at once (e.g. neuroscientists studying the brain at the level of ion channels, individual neurons, local field potentials, or functional magnetic resonance signals). However, causal emergence implies that the macro level with highest *EI* is the one that is optimal to characterize, predict, and retrodict the behavior of the system – the one that "carves nature at its joints" (Hamilton and Cairns, 1961).

The search for the macro level at which *EI* is maximal has a parallel in information theory: channel capacity is an intrinsic property defined as the maximal amount of information that can be transmitted along the channel at a certain rate, found by searching over all possible input distributions (Shannon, 1948). Finding the optimal level of coarse graining for causal emergence is based on a similar search, with several differences. First, *EI* is evaluated using perturbations over the system itself, rather than across a channel (the system is its own input and output). Second, the probability distributions over micro states that can be considered must conform to a proper mapping of micro into macro elements (or time intervals). Additional connections of *CE* to established measures, such as reversibility and lumping in markov processes (Kemeny and Snell, 1976), or epsilon machines (Shalizi and Crutchfield, 2001), are a potential subject for future work.

Causal exclusion and its implications. If causal interactions are strongest at a macro level (causal emergence), multiple causation can only be avoided if causation at the macro level *supersedes* causation at the micro level (*causal exclusion*) (Kim, 1993). Supervenience means that there cannot be additional causal interactions at the macro level that are unaccounted for at the micro level. It is typically concluded that since micro causes are complete (include all details), they do all the causal work. Thus, admitting any macro causation would amount to double-counting causes.

Causal analysis as presented here endorses both supervenience (no extra causal ingredients at the macro level) and causal exclusion (for a given system at a given time, causation occurs at one level only). However, causal analysis also demonstrates that *EI* can actually be maximal at a macro level, depending on the system's architecture. In such

cases, causal exclusion turns the reductionist assumption on its head, since to avoid double-counting causes, optimal macro causation must exclude micro causation. In other words, macro mechanisms can always be decomposed to their constituting micro mechanisms (supervenience); however, if there is emergence, macro causation does not reduce to micro causation, in which case the macro wins causally against the micro and takes its place (supersedence). The notion of irreducibility among levels (does the macro beat the micro?) is complemented by the notion of irreducibility among subsets of elements within a level (is the whole more than its parts? (Albantakis et al., 2013; Tononi, 2012)). From the perspective of a system, emergence ($CE>0$) implies causal "self-definition" at the optimal macro level – the one at which its causal interactions "come into focus" (Alexander, 1920) and "the action happens."

Applicability to real systems. Measuring EI exhaustively, across all micro/macro levels, is not feasible for complex physical or biological systems (see Text S5). Yet, some useful guidelines can be derived from the above analysis: 1.) if $Eff(S_m) \geq Eff(S_m)$ then causal emergence is impossible and causal reduction holds; 2.) if $EI(S_m) > log_2(n_M)$, where n_M is the state repertoire size of S_M, causal reduction holds; 3.) if for some coarse graining Eff increases drastically, causal emergence CE is to be suspected (as $\Delta I_{Eff} >> \Delta I_{Size}$). Therefore systems that already are close to maximal effectiveness at the micro level (Fig. AI.1) indicate causal reduction. By contrast, heavily inter-connected groups of elements with spontaneous activity and the ability to distinguish between intra and inter-group connections, such as the simplified neural system of Fig. 1.6, are more suitable for emergence.

In real neural systems one could compare the value of effective information obtained at the micro scale of single neurons over millisecond intervals, the mesoscale of neuronal groups over hundreds of milliseconds, and the macro scale of brain regions over several seconds (using tools such as optogenetics and calcium imaging). In this way classic notions, such that cortical minicolumns may constitute the fundamental units of brain function (Buxhoeveden and Casanova, 2002), or that the cortex works by population coding in space (Georgopoulos et al., 1986) or rate coding in time (London et al., 2010) in the face of high inter-trial variability (Knoblauch and Palm, 2005), could then be tested rigorously using a measure of effectiveness. Examining small motifs that are overrepresented in complex networks (such as brains (Sporns, 2010)) could determine whether the network as a whole is biased towards emergence or reduction. Heuristic assessments of the likelihood of emergence could also rely on the analysis of wiring diagrams, which can offer an estimate of degeneracy, combined with knowledge of the amount of intrinsic noise in a system, which can provide an estimate of determinism.

Conclusions

The approach to emergence investigated here provides theoretical support for the intuitive idea that, to find out how a system works, one should find the ”differences that make [most of] a difference” to the system itself (Tononi, 2012), cf. (Fitelson and Hitchcock, 2010). It also suggests that complex, multi-level systems such as brains are likely to “work” at a macro level because, in biological systems, selectional processes must deal with unpredictability and lead to degeneracy (Tononi et al., 1999). This may also apply to some engineered systems designed to compensate for noise and degeneracy.

More broadly, this view of causal emergence suggests that the hierarchy of the sciences, from microphysics to macroeconomics, may not just be a matter of convenience but a genuine reflection of causal gains at the relevant levels of organization.

Acknowledgments

We thank M. Boly, C. Cirelli, A. Hashmi, C. Koch, L. Morton, A. Nere, M. Oizumi, and L. Shapiro for helpful discussions, and P. Rana for assisting with the Python software. This work has been supported by the DARPA grant HR 0011-10-C-0052 and the Paul G. Allen Family Foundation.

References

Albantakis L, Hoel E, Tononi G. Intrinsic causation and consciousness. *Association for the Scientific Study of Consciousness* 2013.

Alexander S. *Space, time, and deity: the Gifford lectures at Glasgow, 1916-1918*. 1920.

Bar-Yam Y. A mathematical theory of strong emergence using multiscale variety. *Complexity* 9: 15–24, 2004.

Bedau M. Weak emergence. *Noûs* 31: 375–399, 1997.

Bontly TD. The Supervenience Argument Generalizes. *Philos Stud* 109: 75–96, 2002.

Broad C, Paul T. The mind and its place in nature. In: , edited by Paul T. London: Routledge & Kegan Paul, 1925, p. 97–113.

Butterfield J. Laws, causation and dynamics at different levels. *Interface Focus* 2: 101–114, 2012.

Buxhoeveden D, Casanova M. The minicolumn hypothesis in neuroscience. *Brain* 125: 935–951, 2002.

Chalmers D. Strong and weak emergence. The Reemergence of Emergence, eds Clayton P, Davies P (Oxford Univ Press, Oxford), pp 244–256, 2006.

Davidson D. Mental events. In: *Readings in philosophy of psychology*. 1980, p. 107–119.

Edelman G. *Neural Darwinism: The theory of neuronal group selection.* New York, NY: Basic Books, 1987.

Fitelson B, Hitchcock C. *Probabilistic measures of causal strength*. Oxford University Press, 2010.

Fodor JA. Special sciences (or: The disunity of science as a working hypothesis). *Synthese* 28: 97–115, 1974.

Georgopoulos A, Schwartz A, Kettner R. Neuronal population coding of movement direction. *Science (80-)* 233: 1416–1419, 1986.

Hamilton E, Cairns H. *The collected dialogues of Plato: Including the letters*. Pantheon Books, 1961.

Hölldobler B, Wilson E. *The superorganism: the beauty, elegance, and strangeness of insect societies*. WW Norton & Company, 2009.

Jahr CE, Stevens CF. Voltage dependence of NMDA-activated macroscopic conductances predicted by single-channel kinetics. *J Neurosci* 10: 3178–3182, 1990.

Kemeny J, Snell J. *Finite markov chains*. New York: Springer Verlag, 1976.

Kim J. *Supervenience and Mind: Selected philosophical essays*. Cambridge University Press, 1993.

Kim J. *Mind in a physical world: An essay on the mind-body problem and mental causation*. Cambridge, MA: MIT Press, 2000.

Knoblauch A, Palm G. What is signal and what is noise in the brain? *Biosystems* 79: 83–90, 2005.

Kullback S. *Information theory and statistics*. Dover publications, 1997.

London M, Roth A, Beeren L, Häusser M, Latham P. Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex. *Nature* 466: 123–127, 2010.

Markram H. The blue brain project. *Nat Rev Neurosci* 7: 153–160, 2006.

Pearl J. *Causality: models, reasoning and inference*. 2000.

Sawyer R. *Social emergence: Societies as complex systems*. Cambridge University Press, 2005.

Seth A. Measuring emergence via nonlinear Granger causality. *ALIFE*. 545-552, 2008.

Shalizi C, Crutchfield J. Computational mechanics: Pattern and prediction, structure and simplicity. *J Stat Phys* 104: 817–879, 2001.

Shannon CE. The mathematical theory of communication. 1963. *MD Comput Comput Med Pract* 14: 306–17, 1948.

Sperry R. *Science and moral priority: Merging mind, brain, and human values*. New York: Columbia University Press, 1983.

Sporns O, Tononi G, Kötter R. The human connectome: A structural description of the human brain. *PLoS Comput Biol* 1: e42, 2005.

Sporns O. *Networks of the Brain*. MIT Press, 2010.

Stalnaker R. Varieties of supervenience. *Philos Perspect* 10:221–241, 1996

Tononi G, Sporns O, Edelman GM. Measures of degeneracy and redundancy in biological networks. *Proc Natl Acad Sci U S A* 96: 3257–3262, 1999.

Tononi G, Sporns O. Measuring information integration. *BMC Neurosci* 4: 31, 2003.

Tononi G. Integrated Information Theory of Consciousness: An Updated Account. *Arch Ital Biol* 150: 56–90, 2012.

CHAPTER TWO

Can the macro beat the micro?
Integrated information across spatiotemporal scales

Erik P Hoel[1], Larissa Albantakis[1], William Marshall[1], Giulio Tononi[1*]

[1] Department of Psychiatry, University of Wisconsin, Madison, WI, USA

In preparation.

Abstract

Causal interactions within complex systems such as the brain can be analyzed at multiple spatiotemporal levels. It is widely assumed that the micro level is causally complete, thus excluding causation at the macro level. However, by measuring effective information – how much a system's mechanisms constrain its past and future states – we recently showed that causation can be stronger at macro rather than micro levels. In this work, we go beyond effective information and consider the crucial requirement that, from the intrinsic perspective of a system, a proper measure of causation must also take into account composition (the cause-effect power of the parts), integration (the causal irreducibility of the whole to its parts), and exclusion (the causal borders of the system). A measure satisfying these requirements, called Φ^{Max}, was developed in the context of integrated information theory (IIT), according to which conscious systems are maxima of intrinsic, compositional, irreducible cause-effect power. Here we evaluate Φ^{Max} systematically at micro and macro levels in space and time using simplified neuronal-like systems and show that, for systems characterized by indeterminism and/or degeneracy, Φ can indeed peak at a macro level. This happens if coarse-graining micro-elements produces macro mechanisms having high causal selectivity. These results are relevant to a theoretical account of consciousness, because for IIT the spatiotemporal maximum of integrated information peaks is the spatiotemporal scale of consciousness. More generally, these results show that the notions of macro causal emergence and micro causal exclusion also hold from the intrinsic perspective of a system.

Introduction

The causal structure of physical systems can be analyzed at various spatial or temporal levels, from the most fine-grained micro level to any possible macro coarse-grainings. For example, the brain can be analyzed at the level of neurons, neuronal groups, macro-columns, and areas, over tens, hundreds, or thousands of milliseconds (Sporns et al., 2005). Usually, lack of detailed data, practical considerations, and heuristic strategies dictate the scale at which a system's causal structure can be studied, which is often very coarse-grained. Thus, neuroimaging studies of effective connectivity in the brain examine interactions at the spatial level of voxels, which contain millions of neurons, and at the temporal level of blood oxygen fluctuations, in the order of seconds. While such coarse-grained investigations are useful, it is widely assumed that the causal structure of a system is only fully captured by the most detailed and fine-grained causal model. This 'micro' assumption is ubiquitous in science and underlies ambitious programs that aim at collecting and modeling data at the finest scale possible (Markram, 2006).

At a theoretical level, this reductionist view of causal power is based on the intuition that, since macro-level properties are fixed once the properties of micro-level physical mechanisms are fixed (supervenience), causal power resides fully at the microphysical level (causal closure). Moreover, if all the causal work is done at the micro level, there is no room for any causal contribution at the macro level (causal exclusion; Kim (2000)). Contrary to this common assumption, we recently showed that, once causal interactions are actually measured, causal power at the macro level can surpass that at the micro level (Hoel et al., 2013). This was done by evaluating *effective information*, a

general measure of causal interactions obtained by applying all possible perturbations to capture the how much a system's mechanisms constrain the past and future states of the system as a whole (Tononi and Sporns, 2003). This work demonstrated that effective information can be higher at a macro level if a system is more deterministic (reduced causal divergence) and/or less degenerate (reduced causal divergence) at the macro than at the micro level, counteracting the reduced size of the state space (degrees of freedom).

The analysis of effective information provides a proof of principle that, if causation is evaluated quantitatively for an extrinsically defined system, taken as a whole and on average, 'the macro can beat the micro.' However, from the intrinsic perspective of a system, the assessment of causation should be further refined, to ensure that 'the macro can beat the micro' even when taking into account the cause-effect power of a system's parts (composition), specific system states (specificity), the requirement that the system be irreducible to its parts (integration), and the way the system's borders are defined (exclusion). In the present work, we set out to assess cause-effect power systematically – at all spatiotemporal levels of simple systems - by employing a measure that takes into account these requirements. This measure, called Φ^{Max}, was developed in the context of integrated information theory (IIT), in order to assess maxima of intrinsic, irreducible, compositional cause-effect power, in a specific, state-dependent manner (Tononi, 2012; Oizumi et al., 2014). The measure has already been applied to classify the causal structure of discrete dynamical systems, such as cellular automata, (Albantakis and Tononi, 2015), and to track how the causal structure of simulated organisms, called animats, evolves in a simulated environment (Albantakis et al., 2014).

Measures of integrated information were originally developed with the objective of characterizing which kinds of physical systems can support conscious experience (Tononi, 2004, 2008), and related measures are being applied successfully in empirical studies aimed at identifying the neural substrate of consciousness (Seth, 2008; Seth et al., 2011; Casali et al., 2013). But even if the brain regions necessary and sufficient for experience are identified, one needs to ask why consciousness should occur at the particular spatiotemporal scale it does (Tononi, 2004; Marom, 2010; Chalmers, 2013). Consciousness appears to ’flow’ at a definite temporal scale (tens to hundreds of milliseconds), and the neural correlates of consciousness are often most precisely characterized at the level of neurons or neuronal groups. IIT provides a principled answer to this question: if the substrate of consciousness is a maximum of integrated information, the elements and time intervals constituting such substrate should also be at the spatiotemporal grain that maximize information integration – that is, at the scale at which a system makes most of a difference to itself (Tononi, 2004; Tononi et al., in press). Therefore, demonstrating that a maximum of intrinsic cause effect power can occur at a macro scale (such as neuronal groups over hundreds of milliseconds) is relevant not only for characterizing causal emergence, but also for a theoretical account of the physical substrate of consciousness. What follows shows, using simple, idealized systems, that intrinsic, irreducible, compositional cause-effect power (Φ) can indeed reach a maximum at a macro scale, both in space and in time.

Theory

Integrated Information Theory. A detailed description of IIT 3.0 can be found in Oizumi et al. (2014). In the following, we will outline the IIT 3.0 algorithm to find the set of elements in a system with the maximum amount of integrated information (Φ^{Max}). This algorithm takes the form of a search across all possible sets of elements, while the system is in a particular state. Due to computational constraints we restrict our search to small networks of binary logic gates.

Φ is computed independently for each "candidate set": each member of the power set of elements in the system. Elements outside of the candidate set are considered exogenous elements and fixed in their state throughout the analysis (they are treated as background conditions).

For a given candidate set *S* (which may be the whole system), we first perturb the set *S* into all possible states with equal probability and record the resulting state distribution at the next time step in a transition probability matrix (TPM), from which all IIT measures can be derived. Formally, this is done using the *do*(*x*) operator (Pearl, 2000): $do(S = s_i)$, where *N* is the number of elements in the set *S*.

The next step in calculating Φ is to determine the cause-effect structure of the candidate set *S*. To that end, every "candidate mechanism" (subset of *S*) must be examined with respect to the causal role it plays in the system, represented by its *maximally irreducible cause and effect repertoires*. These probability distributions of possible causes (the cause repertoire) and possible effects (the effect repertoire) are derived from the TPM (conditioned on the mechanism's current state) and tested for irreducibility as outlined next.

Mechanisms can be either single elements (1st-order) or combinations of elements (higher-order). To exist from the intrinsic perspective of the system, a particular mechanism must play an irreducible causal role within the candidate set. This is assessed by the mechanism's integrated information φ ("small phi"). To obtain the φ value of a mechanism *M* in its current state *m*, *M* is partitioned, which means severing connections by replacing them with noise. φ quantifies the difference between the original (unpartitioned) cause or effect repertoire of *M* and the product distribution of the cause or effect repertoires after the partition:

$$\varphi_{Cause}(m, Z_{t-1}) = D\left(p(z_{t-1}|m) \parallel p(z_{t-1}^{(1)}|m^{(1)}) \times p(z_{t-1}^{(2)}|m^{(2)})\right)$$

$$\varphi_{Effect}(m, Z_{t+1}) = D\left(p(z_{t+1}|m) \parallel p\left(z_{t+1}^{(1)}\middle|m^{(1)}\right) \times p\left(z_{t+1}^{(2)}\middle|m^{(2)}\right)\right).$$

Above, $p(z_{t\pm1}|m)$ denote the cause and effect repertoires of the mechanism *M* in state *m*, respectively, over a particular set of elements from the candidate set, the "purview" $Z_{t\pm1}$, where Z_{t-1} can differ from Z_{t+1}. In IIT, distances *D* between probability distributions are measured using the earth-mover's distance (EMD) (vvv IIT 3.0), which quantifies the minimum cost of transforming one probability distribution into another (Rubner et al., 2000; Pele and Werman, 2009). To find φ_{Cause} all possible bipartitions $P = \left\{M^{(1)}, Z_{t-1}^{(1)}; M^{(2)}, Z_{t-1}^{(2)}\right\}$ of *M* and its past purview Z_{t-1} are attempted and the partition that gives the minimum value of φ is selected (the minimum information partition, or *MIP*). To find φ_{Effect}, the same procedure is performed testing all bipartitions $P = \left\{M^{(1)}, Z_{t+1}^{(1)}; M^{(2)}, Z_{t+1}^{(2)}\right\}$. The overall φ value of the mechanism for a particular set of purviews $Z_{t\pm1}$ is the minimum between its φ_{Cause} and φ_{Effect} (in this way,

if a mechanism receives input but gives no effective output, or vise versa, it has $\varphi = 0$, as it plays no causal role in the system).

While a mechanism can have a positive φ value for different sets of purviews, the causal role of the mechanism in its current state is identified by finding the purviews $Z^*_{t\pm1}$ with maximal $\varphi(m, Z_{t\pm1})$. The cause and effect repertoires identified by this procedure are referred to as the *core cause repertoire* and *core effect repertoire*.

If a mechanism in a state plays an irreducible causal role (by having $\varphi^{Max} > 0$), we refer to it, along with its maximally irreducible cause-effect repertoire, as a *concept*. The set of all concepts makes up the cause-effect structure of the candidate set. In this way, the φ analysis reveals the compositional nature of the cause-effect structure, identifying not only elementary mechanisms, but also irreducible higher-order mechanisms, and how all mechanisms are related.

Having obtained the cause-effect structure of a given candidate set, the next step in the IIT algorithm is to use system partitions to determine if the cause-effect structure, as a whole, is irreducible. This is because in a system with a reducible cause-effect structure, some parts of the system will not make a difference to other parts of the system and in that case the system as a whole cannot exist for itself. Bipartitions of the candidate set *S* are accomplished by a cut (replacement with noise) of the unidirectional connections between a subset of *S* and the remaining elements. Integrated (conceptual) information (Φ, "big phi"), measures the irreducibility of a cause-effect structure, by quantifying the differences the system partition makes to all its concepts and their φ values:

$$\Phi(s) = D\big(C(s)\big) \parallel C_P(s))$$

where C is the unpartitioned cause-effect structure of candidate set S in state s, and C_{P} the cause-effect structure after the partition. The distance D between two cause-effect structures is assessed using an extended version of the earth mover's distance. Above, the cost of transforming a probability distribution into another is given by the amount of probability that needs to be shifted, multiplied by the distance it is moved, which here is given by the Hamming distance between the system states. To calculate the EMD between two cause-effect structures, the cost of transforming C into C_{P} is the amount of φ that needs to be shifted, multiplied by the distance between the concepts, which is given by the distance between the cause-effect repertoires of those concepts. For full details and examples on the Φ calculation, see Oizumi et al. (2014).

Over all possible bipartitions $P_{\rightarrow} = \{S_1, S_2\}$ of S the minimum information partition is selected that minimizes Φ.

Once all candidate sets are evaluated, the Φ value of all possible candidate sets are compared to find the subset of elements with the highest Φ value, ('Φ^{Max}'). This is the subset of elements that generates the maximally irreducible cause-effect structure, the *complex.* A complex has intrinsic causal borders and exists for itself. By causal exclusion, complexes cannot overlap, as this would multiply causes and effects without necessity ("multiple causation"). According to IIT, there is an identity between the maximally irreducible cause-effect structure of a complex in its current state and its state of consciousness, its experience.

The causal attributes of integrated information (φ). φ measures the difference that a partition $P = \left\{M^{(1)}, Z_{t\pm1}^{(1)}; M^{(2)}, Z_{t\pm1}^{(2)}\right\}$ of mechanism *M* in state *m* makes to the mechanism's cause and effect repertoires over its purviews $Z_{t\pm1}$. This difference to the cause and effect repertoires can be decomposed into three causally contributing features: (1) the size of the cause or effect repertoire, (2) a change in how selective the mechanism is about its possible causes and effects post-partition and (3) a shift in which states are selected as possible causes and effects post-partition (Fig. AI.1).

The *size* of a repertoire is $\log_2(n)$, where *n* is the number of states of the purview. It reflects the degrees of freedom of potential causes or effects.

The *irreducible selectivity* of a mechanism describes how much the mechanism constrains the past and the future above and beyond its parts. Selectivity of a repertoire can be measured as the difference *D* (EMD) between the cause or effect repertoire $p(z_{t\pm1}|m)$ to the maximum entropy distribution $p(H)$ over all possible states of $Z_{t\pm1}$:

$$selectivity(m, Z_{t\pm1}) = D(p(z_{t\pm1}|m) \parallel p(H)).$$

Selectivity can be normalized by the size, $\log_2(n)$, which is also the distance between a maximum entropy distribution and a perfectly selective distribution ($p = 1$). To capture the *irreducible selectivity* of a mechanism above and beyond its parts, we moreover subtract the distance between the partitioned repertoire $p^{MIP}(z_{t\pm1}|m)$ and $p(H)$. In this way:

$$irreducible\ selectivity(m, Z_{t\pm1}) = (D(p(z_{t\pm1}|m)||\ p(H)) - D(p^{MIP}(z_{t\pm1}|m)||\ p(H)))/size$$

Irreducible selectivity values can range between 0.5 and -0.5, inclusively (as negative values are rare but possible).

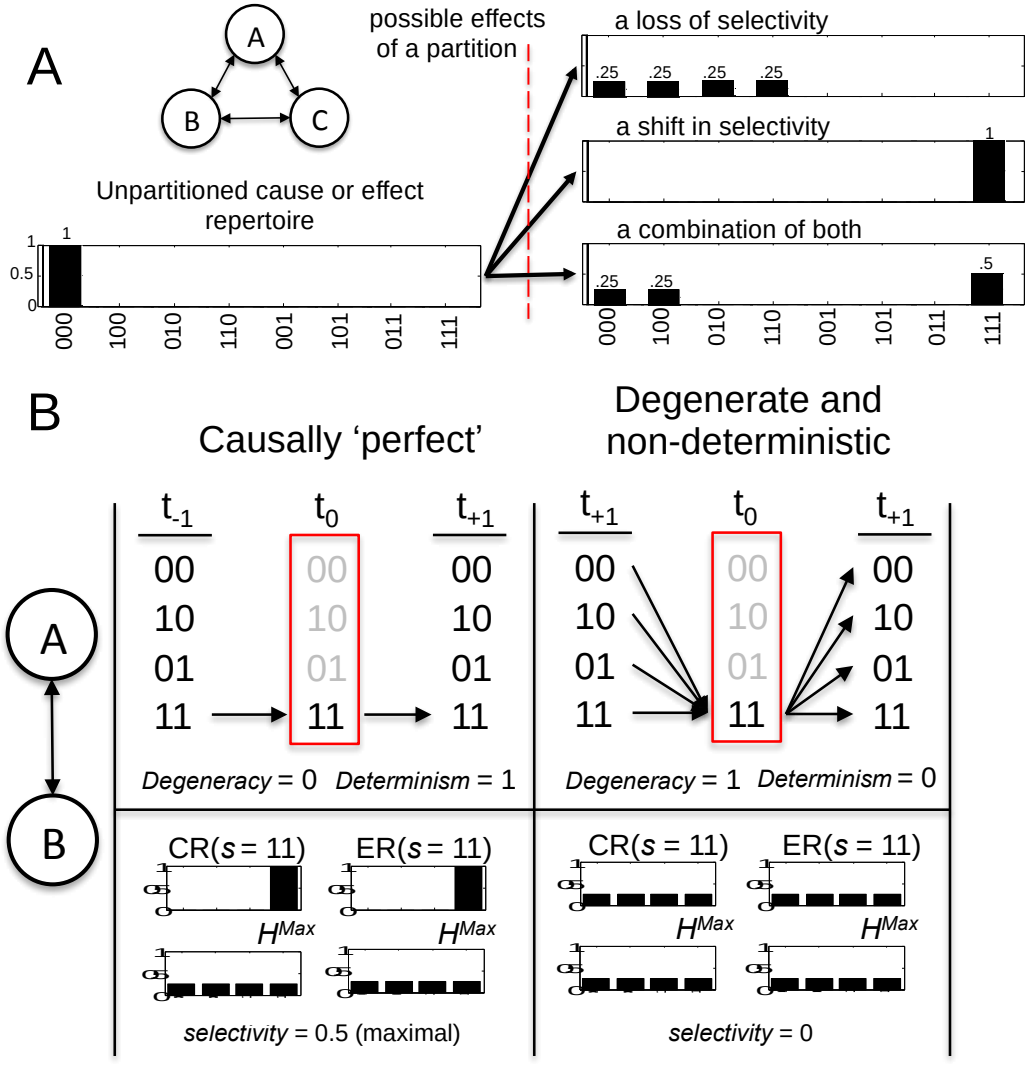


Figure 2.1. *The causal components of IIT (φ).* (A) Consider a hypothetical isolated system constituted of 3 interconnected binary elements. The unpartitioned cause-effect repertoires of the system can change in two ways following a partition. There can be a loss of selectivity, moving the partition closer to maximum entropy (top), a shift in which

states are selected in the partitioned repertoire (middle), or a mix of both (bottom). (B) Consider a simpler system of just 2 connected binary elements (left). If the system in state 11 at t_0 could only originate from 11 at t_{-1}, and can only go to 11 at t_{+1}, then that state has a degeneracy of 0 and a determinism of 1 (B, top left) Compare to if the system in state 11 at t_0 could have originated from any state at t_{-1}, and could go to any state at t_{+1}, all with equal probability (top right), in which case the mechanism in state 11 has a degeneracy of 0 and a determinism of 1. Compare degeneracy and determinism to selectivity: the minimum distance of either the cause or effect repertoires from the maximum entropy distribution (*H*). In both cases, selectivity accurately reflects determinism and degeneracy (bottom).

Selectivity can be related to the notion of determinism and degeneracy (Hoel et al., 2013). Previously, we demonstrated that the effective information in a causal model depends on how collectively deterministic and degenerate its mechanisms are. *Determinism* indicates how reliably the current state of a mechanism leads to future states: determinism is 1 when the current state leads to a single future state with probability $p = 1$, and is 0 when all future states have equal probabilities ($p = 1/n$, where n is the number of states). *Degeneracy* indicates how many states converge to the same state: degeneracy is 1 when the current state could have come from any previous state with probability $p = 1/n$, and is 0 when the current state could only have come from a single previous state with probability $p = 1$. If determinism = 1 and degeneracy = 0 then that mechanism in a state is causally 'perfect' in that it demonstrates maximum selectivity over the states of its purviews (Fig. 2.1B). If determinism = 0 and degeneracy = 1, then there is a total absence of selectivity over the mechanism's purviews (total noise, Fig. 2.1B). Previously we showed that it is through increasing the determinism and/or decreasing the degeneracy of causal relationships that coarse-graining can result in higher cause-effect power (Hoel et al., 2013). Since the selectivity of a mechanism *M* in

state *m* is always positive, it provides the upper limit for the mechanism's irreducible selectivity.

The third causal property captured by φ, the *selectivity-shift*, captures how the mechanism makes a difference to the system by selecting some past or future states over others (see S1 for a detailed explanation of *selectivity-shift*). For example, consider a partition that results in 0 irreducible selectivity, but changes which states are specified by the mechanism (Fig. 2.1A). The resulting non-zero φ value would be due entirely to a selectivity-shift: the rearrangement of probability mass post-partition without a change in the distance from *H*. Shift can be captured by the increase in distance, how much of a detour it is, to pass through the partitioned repertoire on the way from the unpartitioned repertoire to $p(H)$, as opposed to going directly:

$$selectivity\text{-}shift = (D(p(z_{t\pm1}|m)||\, p^{MIP}(z_{t\pm1}|m)) + D(p^{MIP}(z_{t\pm1}|m)||\, p(H)) - D(p(z_{t\pm1}|m)||\, p(H))) \,/\, size.$$

In sum, φ can be fully decomposed into these three quantities: size, irreducible selectivity, and selectivity-shift (full details and proof in Appendix II), as:

$$\varphi = (irreducible\ selectivity + selectivity\text{-}shift) * size$$

Decomposing φ in this manner makes the relevant causal properties of a mechanism apparent.

In IIT, a mechanism's φ value is determined by the minimum of φ_{Cause} and φ_{Effect}. Likewise, the repertoire with minimum φ (φ_{Cause} or φ_{Effect}) determines the mechanism's

size, irreducible selectivity, and selectivity-shift. If the minimum is $\varphi = \varphi_{Effect}$ then irreducible selectivity is bounded by the determinism of the mechanism, whereas if the minimum is $\varphi = \varphi_{Cause}$ then irreducible selectivity is bounded by the mechanism's degeneracy.

φ measures not just the reduction of uncertainty generated by a mechanism *M* in state *m*, but captures the irreducible constraints imposed by *M* = *m*, the differences that make a difference. As a quantity, selectivity-shift describes how much of the difference the mechanism *M* in state *m* makes above and beyond its selectivity, by capturing how much causal work is being done by the mechanism selecting some states instead of others.

The decomposition of a mechanism's φ into the causal properties of size, irreducible selectivity, and selectivity-shift reveals how the different causal properties of each concept uniquely contribute to the overall compositional cause-effect structure of the system.

Coarse-graining systems. A discrete, finite system composed of mechanisms in a state can be considered at various spatiotemporal levels from the most fine-grained micro model of the system (S_m) and at a multitude of coarse-grainings (S_M). When the micro model of a system S_m is fixed, then all its coarse-grainings are also fixed, a property known as "supervenience" (Stalnaker, 1996). At the same time, several different micro-states may all fix the same S_M: a property known as "multiple realizability" (Fodor, 1974).

In this study, our objective is to identify sets of elements that specify global maxima of integrated conceptual information (Φ^{Max}) across elements and spatiotemporal scales. To that end, we extend the algorithmic search for Φ^{Max} across all candidate sets of a system S in micro state s_m to include all sets of coarse-grained ("macro") elements across all spatial and temporal levels of the system in its coarse-grained equivalent macro state s_M.

Herein micro-levels are always composed of binary elements {A, B, C…} with possible states {0,1}. For simplicity, without loss of generality, we confine our analysis to coarse grains in which macro elements are also binary. Macro states will be referred to using {*Off*, *On*} and macro elements using Greek letters {α, β, γ …}. The relationship between the micro-level and any of its macro-levels can be formalized as a mapping, $\mathbb{M}: S_m \rightarrow S_M$, wherein sets of micro elements are grouped into a macro element and the associated micro states are grouped into its macro binary states. In order to be a valid mapping, $\mathbb{M}$ must construct a candidate set S_M wherein the macro elements are conditionally independent (capable of being perturbed independently). Additionally, mappings are limited to those in which the identity of the individual micro elements within a macro element is irrelevant to determine the macro state (or else micro-level information would be available at the macro-level).

To obtain the transition probability matrix (TPM) of the macro level, S_M must be perturbed into all its possible macro states with equal probability, in the same way as done for the micro level. Perturbing a set of macro elements (setting it to a macro state with the *do*(*x*) operator) is done using a macro perturbation, which is the average over

perturbations into all n_{micro} micro states that are grouped into the respective macro state s_M:

$$do(S_M = s_M) = \frac{1}{n_{micro}} \sum_{s_{m,i} \in s_M} do(S_m = s_{m,i})$$

The TPM of a candidate set is thus assessed independently at each respective level, as perturbing S_M into all possible macro states with equal probability typically corresponds to a non-uniform distribution of all possible micro perturbations (except if all macro states are composed of the same number of micro states). This reshaping of micro perturbations at the macro level ultimately allows the "macro to beat the micro", as it makes the causal analysis sensitive to the higher-level causal structure. For more detailed explanations of how system perturbations and coarse-graining is performed herein, see Hoel et al. (2013).

Macro cause-effect structures are calculated from the macro TPM of a macro candidate set as described in the "Integrated information theory" section. This also means that to evaluate the irreducibility φ of the mechanisms within the candidate system at its particular spatio-temporal scale, only partitions between macro elements are permitted, because such partitions reveal the causal structure of the system at a particular level. When calculating Φ of a macro candidate set S_M, however, all partitions possible at the micro level are also performed at the macro level, since Φ measures to what extent S_M exists above and beyond its micro level parts. At the micro and each macro level, the Φ values of all possible candidate sets are evaluated and the candidate set with max(Φ) at each particular level is selected for comparison between levels, yielding the absolute maximum of integrated information (Φ^{Max}) across sets of elements and spatio-temporal scales.

Finally, if Φ^{Max} is found at a macro level, the system demonstrates *causal emergence*. In these cases, max($\Phi(S_M)$) – max($\Phi(S_m)$) indicates how much integrated information is gained by analyzing the system at the macro level. If in the winning mapping macro elements group multiple micro elements along with their micro states at t_0, this is *spatial causal emergence*. If the macro elements consist of only a single micro element but group that element's states over multiple micro timesteps, this is *temporal causal emergence*.

We created all binary coarse grains of discrete systems of logic gates with a custom-made Python program (PyPhi), available for download at www.integratedinformationtheory.org. PyPhi also calculated max(Φ) at each level. Data plots and images were created using MATLAB. Specific examples of spatial and temporal causal emergence are shown below, along with comparisons of macro concepts to their underlying micro concepts.

Results

Finding the maximal value of Φ is an algorithmic search across all possible subsets of a given system. Here, this search is expanded to include all possible binary coarse-grainings. The figures confine themselves to showing only the micro-level value $\Phi(S_m)$, along with the result of the search: the macro level (S_M) with the highest Φ.

IIT analysis of spatial causal emergence. To begin with a simple example, consider a 4-element system S_m = {ABCD} in state [0000], where each micro mechanism operates as an AND gate (with 2 inputs) under noisy conditions (Fig. 2.2A). In Hoel et al. (2013),

this system showed spatial causal emergence in terms of effective information. The results of applying IIT to the micro level S_m can be seen in Fig. 2.2B. Each micro element is associated with a single micro concept, for which $\varphi = 0.17$. To visualize the conceptual structure of the system, each concept is plotted as a star in cause-effect space (Fig. 2.2C). In cause-effect space each dimension is a possible past or future state of the system. Each of the 4 concepts of S_m occupies a position in the 32-dimensional space based on the probability distributions of its maximally irreducible cause-effect repertoires. The size of each star represents how irreducible the concept is (its φ value). Observing the constellation of concepts for the micro-level of the system it is obvious that the concepts are small and clustered together in qualia space, with the concepts of A/B and C/D overlapping. $\Phi(S_m)$ is only 0.11, reflecting the lack of composition and differentiation of the conceptual structure. The *MIP* is the {AC} connections to {BD}.

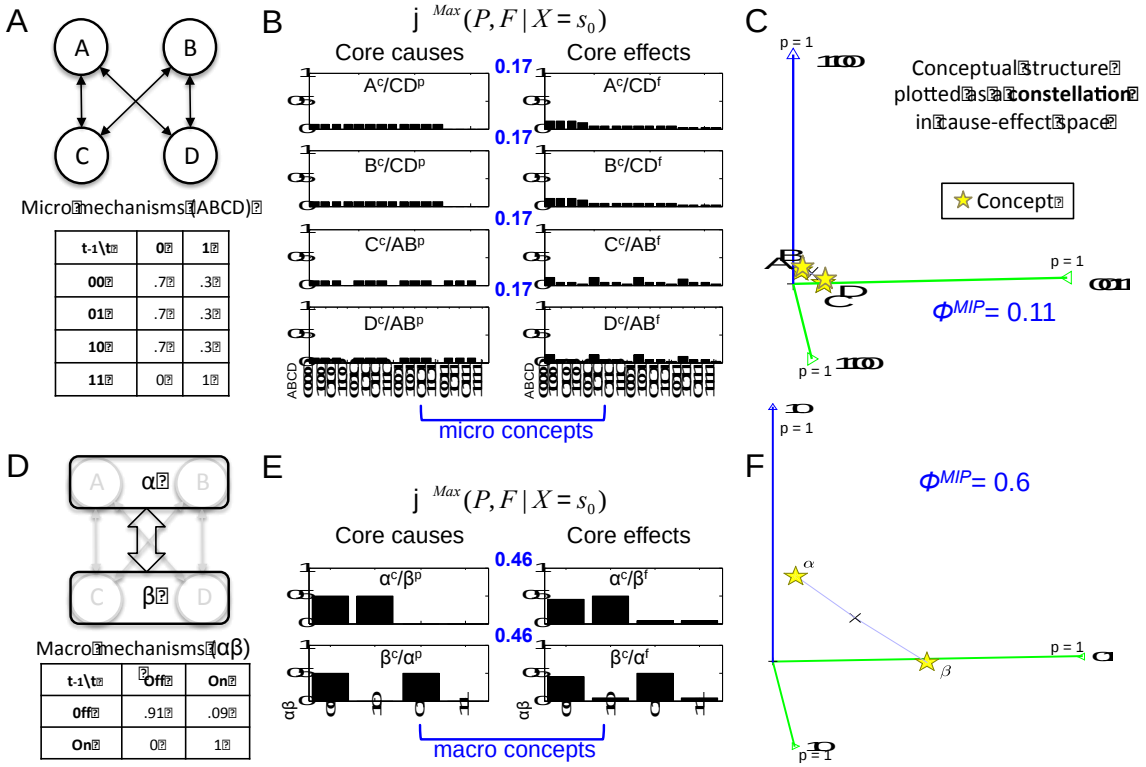


Figure 2.2. *Spatial causal emergence of integrated information (increasing determinism).* (A) The micro level S_m is constituted of noisy mechanisms. (B) The 4 micro concepts all share the same φ value. Shown are the core causes and effects, with the format of A^c/CD^P, which indicates that the concept belongs to A in its current state (c) and has a purview of CD in their past states (p). (C) A 3D projection of the 32-dimensional cause-effect space. The 1 past (blue) and 2 future (green) dimensions chosen were those with the greatest variation of probabilities (so the visualization maximizes the distances between concepts). The conceptual structure of S_m appears as a clustered constellation of small (low φ) concepts. (D) The mechanisms at the macro level of the system S_M are less noisy than those in S_m. (E) The 2 macro mechanisms each generate a concept. (F). The conceptual structure of the macro level system, plotted as a constellation in S_M's 8-dimensional cause-effect space, has larger (high φ) concepts, indicating greater irreducibility.

In contrast, consider the macro level $S_M = \{\alpha,\beta\}$, shown in Fig. 2.2D. The micro-to-macro element mapping is of {AB} to {α} and {CD} to {β}, while the state mapping is such that within each macro element the micro states [00, 01, 10] are considered "Off" and [11] is considered "On." This mapping creates the macro mechanism tables seen at

the bottom of 2D. The macro elements are each associated with a macro concept, for which φ = 0.46 (Fig. 2.2E). The constellation of concepts of S_M in cause-effect space (Fig. 2.2F) In the macro conceptual structure the 2 macro level concepts are much larger (more irreducible) and are less clustered than those in S_m. The average distance (taking pairwise EMDs between the cause-effect repertoires) between all the macro concepts = 1.91, while the average distance between all the micro concepts = 0.9. Reflecting this, Φ^{Max} is at the macro level ($\Phi^{Max}(S_M)$ = 0.6), for the cut of the connections from {BC} to {AD}, not at the micro level.

How does the macro beat the micro*?* Broadly, the macro beats the micro by increasing the causal power (as measured by φ) of redundant or noisy micro mechanisms by grouping them together. This can be quantified directly: as shown in the Theory section, φ can be decomposed into the causal properties of size, irreducible selectivity, and selectivity-shift. Here, we show that while size always decreases with coarse-graining, both irreducible selectivity and selectivity-shift can increase to a degree that outweighs the loss in size, and this allows the macro to beat the micro. Fig. 2.3 shows an example of the unpartitioned and partitioned cause-effect repertoires of {A}, as well as of its supervening macro concept {α}. The size of the micro concept's repertoire = 2, while the size of the macro concept's repertoire = 1. However, at the micro level, both the unpartitioned and partitioned distributions are very close to maximum entropy (shown in blue) and thus the irreducible selectivity of the micro concept is only 0.09. By comparison, the irreducible selectivity of the macro concept is 0.37. The selectivity-shift also changes at the macro level, going from 0 at the micro level to 0.09 at the macro

level. Thus, while the macro loses 0.09 in φ from the loss in size, it gains 0.28 in φ from the increase in irreducible selectivity and 0.09 in φ from the increase in selectivity-shift. Note that the majority of the gain is driven by an increase in irreducible selectivity, not the selectivity-shift. This gain in irreducible selectivity comes from an increase in the determinism (for the macro concept $\varphi = \varphi_{Effect}$, its selectivity is over the future). This is in line with previous results from Hoel et al. (2013) where it was demonstrated that macro level causal relationships could have greater determinism and less degeneracy than their underlying micro level causal relationships. As concepts with higher φ generally contribute more to Φ, systems with higher sums of φ generally have higher levels of Φ (Albantakis et al., 2014; Albantakis and Tononi, 2015). So the greater φ of the macro concepts allows the macro to beat the micro in terms of Φ as well.

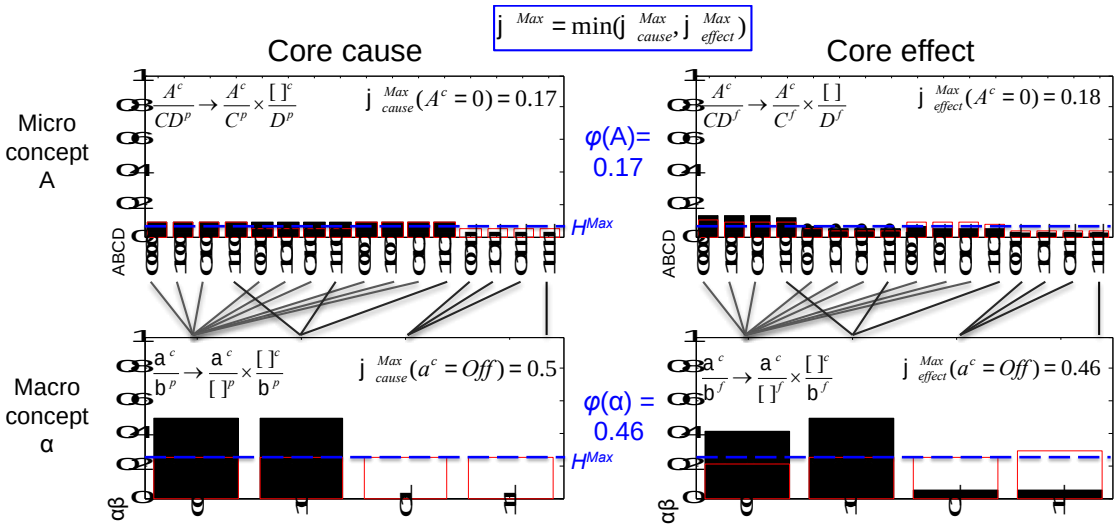


Figure 2.3. *How the macro beats the micro.* A comparison of a micro concept of Fig. 2.2 to its supervening macro concept. (A) The core cause-effect repertoires of element {A} which have been expanded over the whole system, with their respective *MIP* shown in the upper left corners. The unpartitioned repertoires are in solid black, while partitioned repertoires are in red. The dotted blue line shows where the maximum entropy distribution lies. (B) The expanded core cause-effect repertoires of element {α}, which supervenes on {A}.

IIT analysis of spatial causal emergence in a highly degenerate system. Macro concepts can also have higher φ by having less degeneracy than their underlying micro concepts. Consider the micro level of the system shown in Fig. 2.4A. The S_m elements {A-F} are a cycle of deterministic AND gates in state [000000]. As AND gates in the [0] state are highly degenerate (see mechanism table), each concept has φ = 0.167, despite having a repertoire size of 2. The low irreducible selectivity of the concepts (0.083) reflects this degeneracy. The shift for all micro concepts is 0. In cause-effect space the micro concepts are clustered and highly reducible (Fig. 2.4B). $\Phi(S_m)$ = 0.19 for cutting the unidirectional connections from {A} to {BCDEF}. S_M = {αβγ}, where the pairs of {AB}, {CD}, {EF}, are grouped into {α}, {β}, {γ}, respectively. In each macro element the micro states [00, 01, 10] are considered "Off" and [11] is considered "On." Thus the cycle of AND gates at the micro level has been turned into a cycle of COPY gates at the macro level (Fig. 2.4C). At the macro level the *MIP* cuts the connections from {ABCDE} to {F}. $\Phi^{Max}(S_M)$ = 0.83 (Fig 4D). For all macro concepts φ = 0.5. The average distance between the macro concepts = 2, while the average distance between all the micro concepts = 1.33. While the size of the macro concepts' repertoires is reduced to 1, their irreducible selectivity has increased to 1 (their selectivity-shift = 0), meaning that the macro elements have 0 degeneracy. It is this reduction in degeneracy that allows the macro to beat the micro.

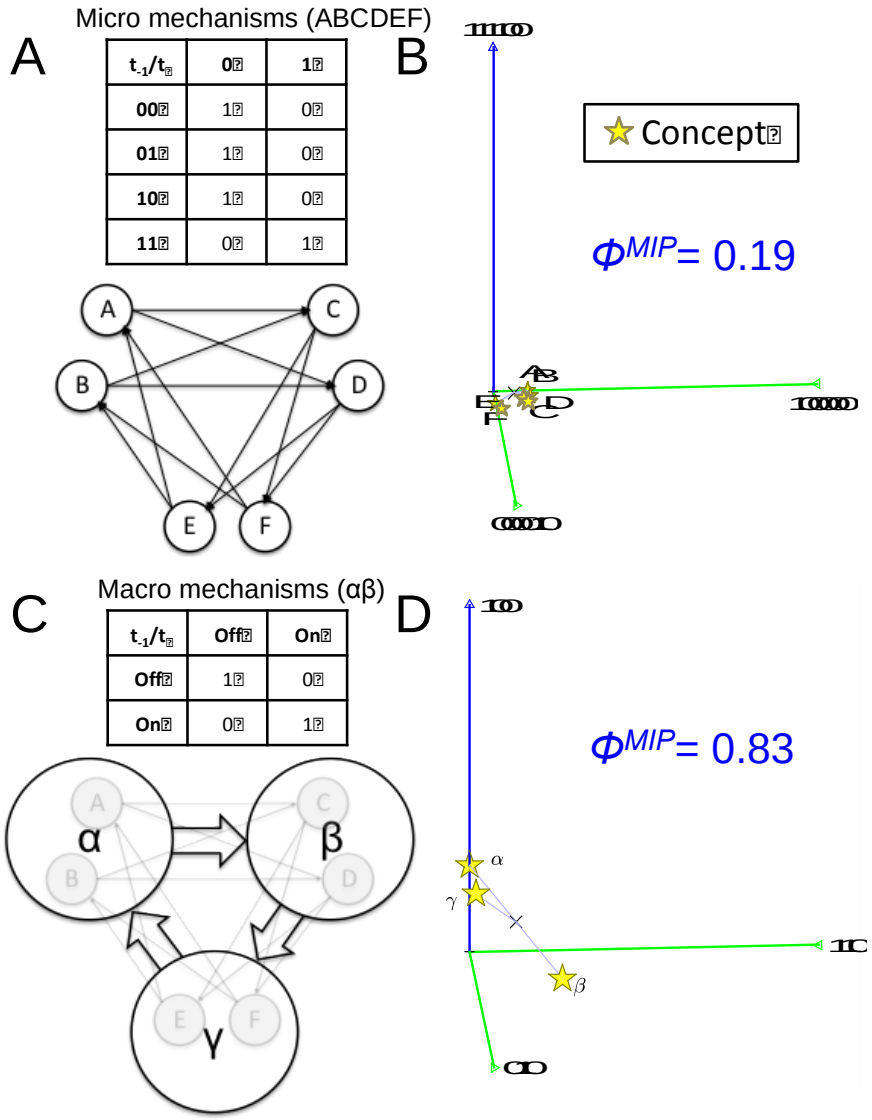
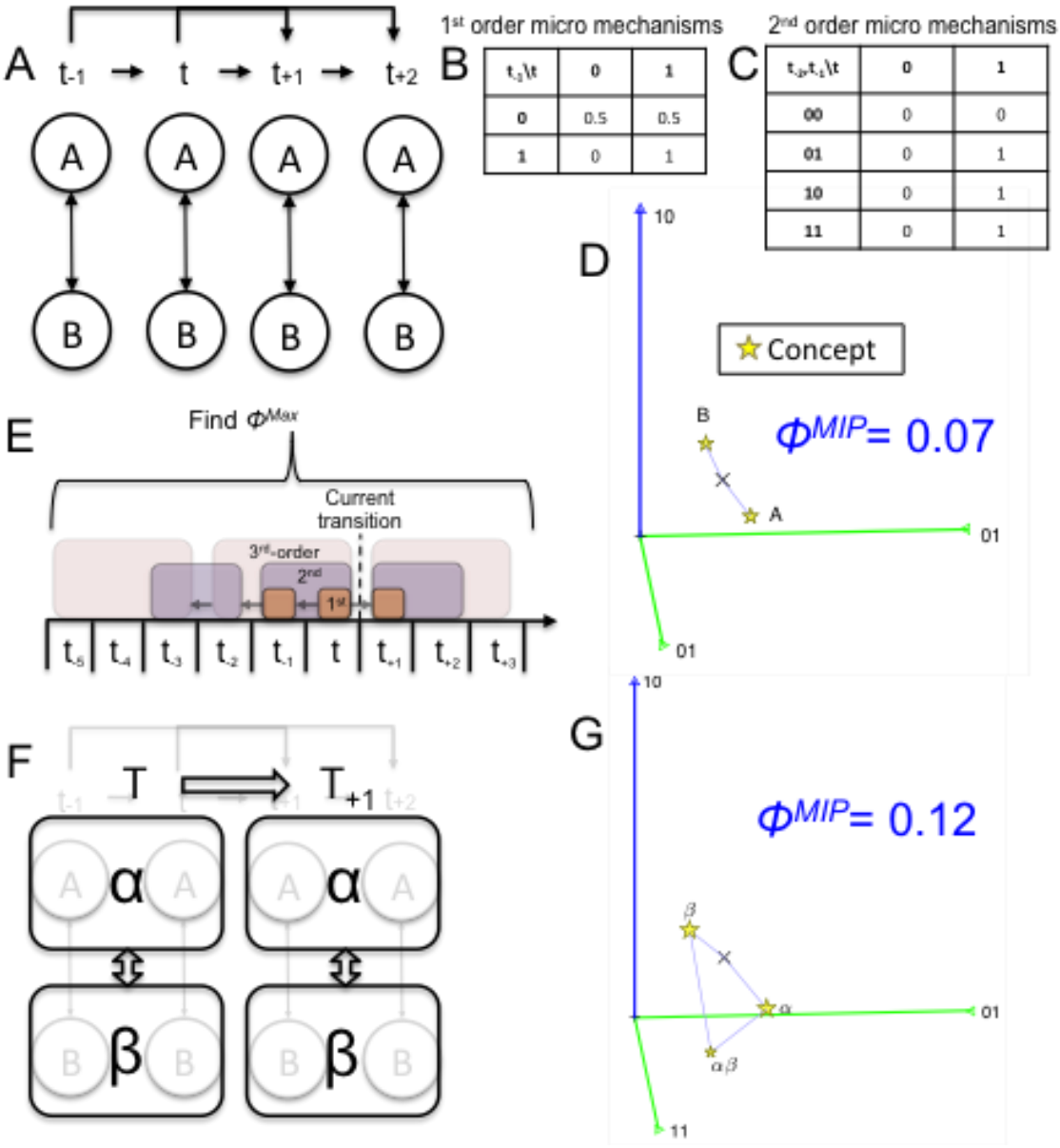


Figure 2.4. *Spatial causal emergence through degeneracy.* (A) A highly degenerate but deterministic S_m is constituted of AND gates. (B) The conceptual structure is highly clustered and the concepts highly reducible. (C) S_M is still deterministic but is no longer degenerate. (D) The conceptual structure is less clustered and less reducible at the macro level.

IIT analysis of temporal causal emergence. Macro groupings may be over time, as well as space (Hoel et al., 2013). In such cases it is the micro timesteps (t_x) that are coarse-grained into macro timesteps (T_x). The system in Fig. 2.5A, {AB} is, from its intrinsic perspective, operating at a particular timescale. This intrinsic timescale is whichever one,

across all possible timescales and all possible coarse-grainings of those timescales, gives Φ^{Max}. For example, at one timestep t_0 the state of the system AB = [11], and the mechanisms are 1st-order (see table in Fig. 2.5B). The system can also be analyzed over 2 timesteps, so that its current state is defined over (t_{-1}, t_0), here AB = [[11][11]] (mechanisms are 2nd-order in Fig. 2.5C). Analyzing {AB} as over one timestep leads to a Φ^{MIP} of 0.07 (Fig.5D), compared to analyzing {AB} over two timesteps, for which Φ^{MIP} is only 0.01. Fig.5E outlines how each possible timescale is analyzed. Each timescale is also coarse-grained in all possible ways in the search for Φ^{Max}. Fig. 2.5C shows one such coarse-graining wherein {A_{t-1}, A_{t0}} are grouped together into a single macro element over a single macro timestep {α_{T0}}, with {B_{t-1}, B_{t0}} grouped into {β_{T0}}. In the mapping, the micro state [00] is considered "Off," while the micro states [01, 10, 11] are considered "On" for each macro element. $\Phi^{Max}(S_M)$ = 0.12, and thus the macro intrinsic timescale the system operates on causally is longer than the most fine-grained timesteps. As the macro is grouping over the two timesteps, it still suffers a loss of the size of its repertoires compared to the micro over two timesteps (going from an average concept repertoire size of 2.5 to 1.33). Average macro irreducible selectivity = 0.16, average selectivity-shift = 0.03; over one timestep the micro level average irreducible selectivity = - 0.125, average shift = 0.25. The average distance for the macro concepts = 0.75; average distance between for the micro concepts = 0.58. The macro conceptual structure in Fig. 2.5G has more concepts than the one timestep micro level shown in Fig. 2.5D. This is because, although at the micro level the one timestep analysis has the highest Φ, the macro with Φ^{Max} is a coarse-grain of the two timestep micro level that has more concepts.



t_{-1}\t	0	1
0	0.5	0.5
1	0	1

$t_{-2}t_{-1}$\t	0	1
00	0	0
01	0	1
10	0	1
11	0	1

Figure 2.5. *Temporal causal emergence.* (A) S_m is in a particular state t. (B) From t, it can be causally analyzed as a 1st-order system, or (C) as a second-order system with an extended mechanism table. (D) Of the timescales, the smallest micro timescale gives the highest Φ^{MIP}. (E) The search for Φ^{Max} across all timescales and coarse-grainings of those timescales, all centered around time t. 1st, 2nd, 3rd, and so forth, mechanisms are considered (represented as different colors), and also coarse-grained. (F) The resulting S_M of this search, a macro system of two elements {αβ} over a single coarse-grained macro timestep (T). (G) The macro conceptual structure, which is less reducible than the underlying, single-timestep micro, and also has an additional concept.

Complexes at different spatial and temporal scales. A complex is the set of elements that specifies a maximum of Φ. The way a system condenses into complexes is different across spatiotemporal grains. For example, the system in Fig. 2.6A, if analyzed at the micro level in state [000000] forms a complex over the full set of elements {A-F}, with a total of 8 micro concepts (Fig. 2.6B). Nevertheless, the Φ^{MIP} is only 0.17 since, despite the average size of their repertoires being 2.25, the concepts have low irreducible selectivity (0.07), demonstrate little selectivity-shift (0.04), and are not very distant (average = 1.08). The *MIP* is the unidirectional cut of the connections from {EF} to {ABCD}. However, extending the search for Φ^{Max} across all possible coarse-grainings, found that the complex with $\Phi^{Max}(S_M) = 0.6$ consists only of the macro elements {αβ}, which are groupings of micro elements {ABCD}, but does not include {EF}. The state grouping is the same as in Fig. 2.2, with the same size (1), irreducible-selectivity (0.37) and selectivity-shift (0.09), as well as average distance (1.91). According to IIT, only the maximum of Φ over both elements and levels qualifies as a complex (other elements and levels are excluded). Therefore, from the intrinsic perspective a system "self-defines" both its set of elements and its spatiotemporal level.

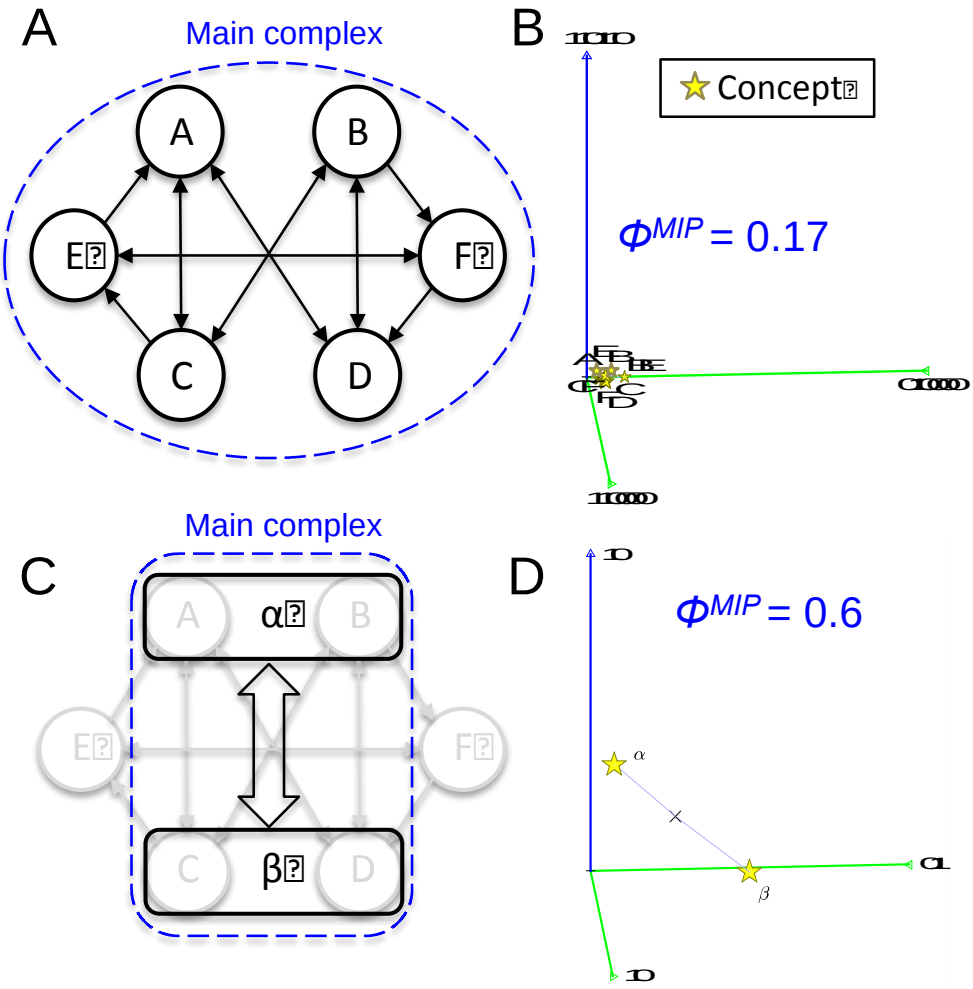


Figure 2.6. *Macro complexes versus micro complexes.* (A) S_m is constituted of heterogeneous logic gates: {ABCD} act identically to those described in Fig. 2.2., while {EF} each act as deterministic AND gates. Additionally, if {E} = 1 at t_{-1}, then the probability of {A} = 1 at t increases by 0.2; the same rule applies for the connection from {F} to {D}. (B) In state [000000], S_m forms a complex {A-F} with 8 concepts. (C) The complex at S_M only supervenes on a subset of the system {ABCD}. (D) The macro conceptual structure has $\Phi^{Max} = 0.6$.

Causal emergence represents maxima in space and time. Φ^{Max} represents the absolute maximum over all different values of Φ^{MIP} found at each possible spatial and temporal scale. In Fig. 2.7, we show the respective Φ^{MIP} values at each possible spatial and temporal level for the four example systems. The winning macro levels appear as clear

maxima for all of the systems. We next examined the relationships between each level of the system. That is, we identified coarse-grains of coarse-grains. For example, consider a micro level {ABCD}, for which the micro elements {AB} are grouped into a macro element {α} with a state mapping of [00, 01, 10] into "Off" and [11] into "On." A further coarse-grain might group {CD} into {β} with the same state grouping, and then the next might group {αβ} together into a single macro element. This nesting of coarse-grains reveals paths of coarse-graining that span all the spatiotemporal levels, from the micro to the macro. Intriguingly, the path from the micro level to the coarse-grain with Φ^{Max} contained the maximum Φ^{MIP} at each level in nearly all the systems. Additionally, instead of being distributed randomly, Φ increased closer to the level where Φ^{Max} was located. Together this suggests that it may be possible to construct heuristics for finding the intrinsic spatiotemporal scale of more complex systems.

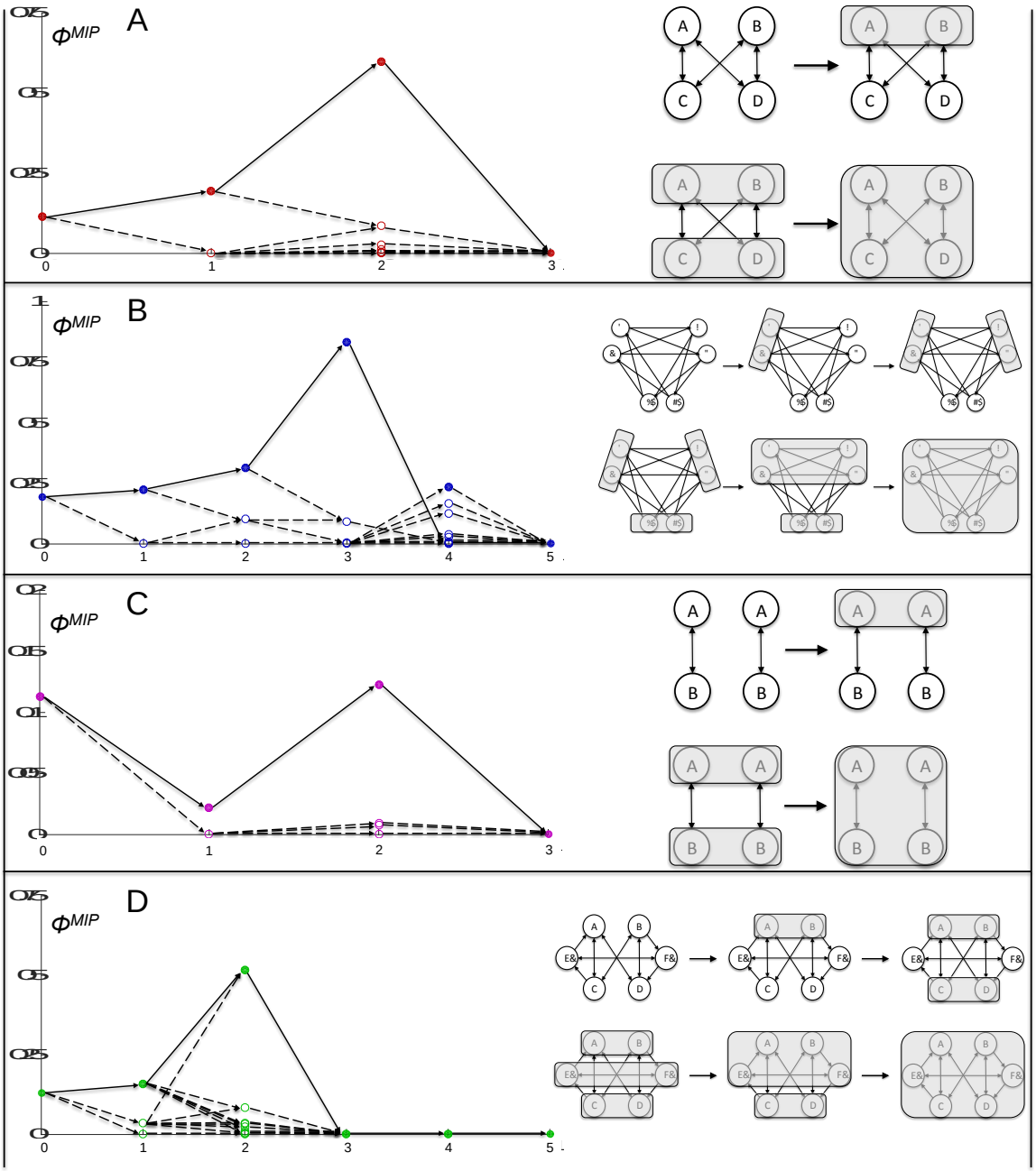


Figure 2.7. *Finding spatial and temporal maxima.* For each of the 4 systems previously examined, the coarse grains at each possible spatiotemporal level are plotted against Φ^{MIP} values (y-axes) on the left of the figure. The x-axes represent the levels of the system: level 0 being the micro level, level 1 being a single grouping of 2 micro elements and so on as the degree of coarse-graining increases until all the elements have been grouped into one macro element (which always has a Φ^{MIP} of 0). The solid color data points represent the maximum Φ^{MIP} value of the groupings of that particular level. The relationships between levels are shown as arrows: each represents a grouping of a lower level. Note that in each example system the path from the micro level to the grouping with Φ^{Max} (tracked by solid arrows) includes the maximum Φ^{MIP} value at nearly every

level. (A) All possible spatial groupings of the 4 element system from Fig. 2.2. (B) All possible spatial groupings of the 6 element system from Fig. 2.4. (C) All possible temporal groupings for the 2 element system in Fig. 2.5. (D) All possible spatial groupings for the 6 element system in Fig. 2.6.

Discussion

Previous work on causal emergence demonstrated that the effective information (the ability to constrain past and future states) of a system can be greater at a macro level than at the micro level (Hoel et al., 2013). This paper extends that work by considering a measure of intrinsic causal power (integrated information, Φ). There are two key differences between effective information and integrated information: Φ considers the compositional structure of the system and it is sensitive to how irreducible a causal structure is (Tononi and Sporns, 2003; Balduzzi and Tononi, 2009; Oizumi et al., 2014). We presented several examples where the integrated information is greater at the macro scale, either spatially or temporally, in systems that are both deterministic and stochastic. These results demonstrate that, from the intrinsic perspective of the system there can be causal emergence of macro levels and exclusion of micro levels.

Causal emergence contradicts the reductionist "exclusion argument" that the micro level of a system causally excludes all macro levels (Kim, 2000).While from an extrinsic perspective a system can be causally modeled at any level useful for an agent, from the intrinsic perspective that system has a particular spatial and temporal scale based on where Φ reaches a maximum. Therefore, we've shown that from the intrinsic perspective, it is possible that the macro level is rather the micro level is causally excluded. This is because the macro can have both more causal power and also be less reducible than the micro.

To calculate Φ across all levels is computationally demanding even for the small example systems considered here; in order to perform this analysis for larger systems will require some smart search algorithm. For each example presented in this work, there is a "path" of coarse-grained systems with the highest Φ values at each level that eventually leads to the maximum value of Φ (Figure 7). This result suggests an algorithm for searching spatial and temporal levels that could greatly speed up computational time, as well the possibility of heuristics which could be applied in real systems, such as the brain.

How the macro beats the micro. Additionally, herein we showed that irreducible causal relationships can be decomposed into three causal properties: size, irreducible selectivity, and selectivity-shift. The size of a causal relationship is the degrees of freedom of the cause-effect repertoire; it is reduced in the process of coarse-graining macro elements, leading to a reduction in causal power. This means that all coarse-grains are at a disadvantage in terms of causal power due to their smaller size. Irreducible selectivity is how much a mechanism constrains the past or the future states of its purview, above and beyond its parts, and selectivity-shift is how much information the mechanism generates by being over different states as a whole than as its parts. Unlike size, irreducible selectivity and selectivity-shift have the potential to increase causal power as a result of coarse-graining elements. Causal emergence occurs when the macro level has greater causal power than the micro level, i.e., when the increase in causal power due to irreducible selectivity and selectivity-shift outweighs the loss of causal power from reduced size. In the examples presented above, and in line with Hoel et al. (2013) it is an

increase in irreducible selectivity, rather than selectivity shift, which is the dominant factor for explaining why the macro beats the micro.

Can causal emergence occur beyond the class of discrete systems shown here, in actual physical systems? As discussed in Hoel et al. (2013), if actual physical systems at the microscopic level of events are "causally perfect" (zero degeneracy and complete determinism) then they cannot causally emerge by increasing selectivity. The example of Figure 5 provides an inkling that there may be other ways for the macro to beat the micro, e.g., by increasing the number of joint causes and effects in the system. Future work will explore alternative methods of creating macro elements, for example, by treating them as "black boxes." By expanding the definition of macro elements and exploiting these different avenues for causal emergence, it may be that causal emergence is more common in physical systems than has been thought.

Consciousness. According to IIT, the physical substrate of consciousness (PSC) is the set of elements at a particular spatiotemporal grain which supports conscious experience (Tononi et al, in prep). Neuroscientists have differing opinions about what the relevant scales are; whether they are neurons or groups of neurons, if all neurons matter or only specific neuronal populations, if individual spikes count or only synchronized local field potentials, etc. Current empirical evidence suggests that the temporal scale of consciousness on the order of 100-1000ms range (Bachmann, 2000; Holcombe, 2009) and that the relevant elements are on the order of neurons or possibly even cortical minicolumns (Buxhoeveden and Casanova, 2002). Currently, this presents a problem because the spatiotemporal scale of measurement chosen by the observer drastically

affects whether an event is information rich or barren (Panzeri et al., 2010). Our theoretical work presents a solution to this problem, providing a framework for finding the set of elements and spatiotemporal scale for which the brain's capacity for integrating information peaks, and thus also the PSC according to IIT.

Our work provides a framework for testing this prediction of IIT, by assessing the integrated information of the brain across spatiotemporal scales. If the predictions are correct the next step is to test the theory in real or physiologically-realistic model systems. For example, as there is some evidence that different species may have a different temporal scale at which they integrated sensory information (Healy et al., 2013), the analysis outlined herein could reveal whether their PSC is at a different temporal scale as well.

References

Albantakis L, Hintze A, Koch C, Adami C, Tononi G. Evolution of Integrated Causal Structures in Animats Exposed to Environments of Increasing Complexity. *PLoS Comput Biol* 10: e1003966, 2014.

Albantakis L, Tononi G. The Intrinsic Cause-Effect Power of Discrete Dynamical Systems—From Elementary Cellular Automata to Adapting Animats. *Entropy* 17: 5472–5502, 2015.

Bachmann T. *Microgenetic approach to the conscious mind*. John Benjamins Publishing, 2000.

Balduzzi D, Tononi G. Qualia: The geometry of integrated information. *PLoS Comput Biol* 5, 2009.

Buxhoeveden D, Casanova M. The minicolumn hypothesis in neuroscience. *Brain* 125: 935–951, 2002.

Casali AG, Gosseries O, Rosanova M, Boly M, Sarasso S, Casali KR, Casarotto S, Bruno M-A, Laureys S, Tononi G, Massimini M. A theoretically based index of consciousness independent of sensory processing and behavior. *Sci Transl Med* 5: 198ra105, 2013.

Chalmers D. The combination problem for panpsychism. *Panspsychism Reef*. 2013.

Fodor JA. Special sciences (or: The disunity of science as a working hypothesis). *Synthese* 28: 97–115, 1974.

Healy K, McNally L, Ruxton GD, Cooper N, Jackson AL. Metabolic rate and body size are linked with perception of temporal information. *Anim Behav* 86: 685–696, 2013.

Hoel EP, Albantakis L, Tononi G. Quantifying causal emergence shows that macro can beat micro. *Proc Natl Acad Sci U S A* 110: 19790–5, 2013.

Holcombe AO. Seeing slow and seeing fast: two limits on perception. *Trends Cogn Sci* 13: 216–21, 2009.

Kim J. *Mind in a physical world: An essay on the mind-body problem and mental causation*. Cambridge, MA: MIT Press, 2000.

Markram H. The blue brain project. *Nat Rev Neurosci* 7: 153–160, 2006.

Marom S. Neural timescales or lack thereof. *Prog Neurobiol* 90: 16–28, 2010.

Oizumi M, Albantakis L, Tononi G. From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Comput Biol* 10: e1003588, 2014.

Panzeri S, Brunel N, Logothetis NK, Kayser C. Sensory neural codes using multiplexed temporal scales. *Trends Neurosci* 33: 111–20, 2010.

Pearl J. *Causality: models, reasoning and inference*. 2000.

Pele O, Werman M. Fast and robust earth mover's distances. *Proc. IEEE Int. Conf. Comput. Vis.* (2009). doi: 10.1109/ICCV.2009.5459199.

Rubner Y, Tomasi C, Guibas LJ. Earth mover's distance as a metric for image retrieval. *Int J Comput Vis* 40: 99–121, 2000.

Seth A. Measuring emergence via nonlinear Granger causality. *ALIFE*. 545-552, 2008.

Seth AK, Barrett AB, Barnett L. Causal density and integrated information as measures of conscious level. *Philos Trans A Math Phys Eng Sci* 369: 3748–67, 2011.

Sporns O, Tononi G, Kötter R. The human connectome: A structural description of the human brain. *PLoS Comput Biol* 1: e42, 2005.

Stalnaker R. Varieties of supervenience. *Philos Perspect* 10:221–241, 1996

Tononi G, Sporns O. Measuring information integration. *BMC Neurosci* 4: 31, 2003.

Tononi G. An information integration theory of consciousness. *BMC Neurosci* 5: 42, 2004.

Tononi G. Consciousness as integrated information: a provisional manifesto. *Biol Bull* 215: 216–242, 2008.

Tononi G. Integrated Information Theory of Consciousness: An Updated Account. *Arch Ital Biol* 150: 56–90, 2012.

CHAPTER THREE

Synaptic refinement during development and its effect on slow wave activity – a computational study

Erik P Hoel[1], Larissa Albantakis[1], Chiara Cirelli[1], Giulio Tononi[1]
[1] Department of Psychiatry, University of Wisconsin, Madison, WI, USA

Under review at:
The Journal of Neurophysiology

Abstract

Recent evidence suggests that synaptic refinement, the reorganization of synapses and connections without significant change in their number or strength, is important for the development of the visual system of juvenile rodents. Other evidence in rodents and humans shows that there is a marked drop in sleep slow wave activity (SWA) during adolescence. Slow waves reflect synchronous transitions of neuronal populations between active and inactive states, and the amount of SWA is influenced by the connection strength and organization of cortical neurons. Here, we investigated if synaptic refinement could account for the observed developmental drop in SWA. To this end, we employed a large-scale neural model of primary visual cortex and sections of the thalamus, capable of producing realistic slow waves. In this model, we reorganized intralaminar connections according to experimental data on synaptic refinement: pre-refinement, local connections between neurons were homogenous; post-refinement, neurons connected preferentially to neurons with similar receptive fields and preferred orientations. Synaptic refinement led to a drop in the model's SWA and to changes in slow wave morphology consistent with experimental data. To test whether learning can induce synaptic refinement, intralaminar connections were equipped with spike-timing dependent plasticity (STDP). Oriented stimuli were presented during a learning period, followed by homeostatic synaptic renormalization. This led to activity-dependent refinement accompanied again by a decline in SWA. Together, these modeling results show that synaptic refinement can account for the developmental changes in SWA. Thus, sleep SWA may be used to track non-invasively the reorganization of cortical connections during development.

Introduction

Cortical circuits are formed over successive stages of development in the brain (Espinosa and Stryker 2012) that underlie ongoing cognitive maturation (Craik and Bialystok 2006). An early and critical developmental phase is the formation of selective connectivity among neurons that receive related sensory inputs (Yoshimura et al. 2005; White and Fitzpatrick 2007; Ko et al. 2011). An important question is the extent to which these selective circuits are formed through synaptic pruning - a net reduction in synapses and connections, or by synaptic refinement - the selective reorganization of an initially homogenous connectivity without a change in the net number of synapses (Innocenti and Price 2005; Tau and Peterson 2010).

The primary visual cortex (V1) is a model region for studying developmental changes in cortical connectivity. In the adult V1 lateral axons are clustered, connecting to neurons with similar receptive fields (RFs) and preferred orientations (POs) (Gilbert and Wiesel 1983). The reorganization of lateral cortical connections into such a selective pattern, from an initially homogenous state, is thought to occur early on in development (Callaway and Katz 1991). Recent experiments on synaptic refinement in mice have revealed that, after the establishment of feedforward connections, lateral connections reorganize through two processes that appear to balance each other, so that the overall connectivity rate stays constant (Ko et al. 2013). In one process, neurons with similar visual responses become preferentially connected to each other, a phenomenon that occurs largely independent of visual experience; in the other, neurons that do not respond reliably to visual stimuli become less interconnected, a process that does not occur with

dark rearing (Ko et al. 2013, 2014). Additionally, there is recent evidence that synaptic refinement occurs not just through the formation and elimination of synaptic connections, but also through changes in the strengths of existing synapses. Thus, mouse V1 neurons with similar RFs and POs are more likely to have stronger excitatory connections (Cossell et al. 2015).

Once adolescence starts in humans and adult-like cognitive capacities appear, the brain's metabolism falls off sharply, along with its ability to recover function after injury (Feinberg 1983; Spear 2000). Over this period from pre- to post-adolescence, the total number of synapses as well as the corresponding gray matter volume is thought to follow an inverted U-shaped curve, peaking at pre-adolescence and then decreasing over the course of adolescence (Blakemore and Choudhury; Lenroot and Giedd 2006). Intriguingly, a similar inverted U-shaped curve has also been observed in slow wave activity (SWA) during NREM sleep. NREM SWA (defined as the EEG power in the 0.5 to 4 Hz range) also peaks at the end of childhood and then decreases over the course of adolescence (Kurth et al. 2010). It has been suggested that the decline in the total number of synapses may be the cause of the observed decrease in SWA (Feinberg 1983; Campbell and Feinberg 2009). This hypothesis is supported by computer models showing that the amount of SWA correlates with overall synaptic strength (Esser et al. 2007). However, there is actually little direct experimental evidence that significant synaptic pruning occurs in human adolescents at the time of the decline of SWA (Paus et al. 2008; Petanjek et al. 2011).

A decrease of SWA from pre to post adolescence was also observed in mice (de Vivo et al. 2014) and rats, in which SWA follows an inverted U-shaped curve similar to

the one in humans (Olini et al. 2013). In mice, however, it is clear that most of synaptic pruning occurs long before the developmental drop in SWA, and over the period from pre to post adolescence there is no substantial change in the number of synapses in the whole brain (De Felipe et al. 1997; de Vivo et al. 2014). On the other hand, during the same period there is major synaptic refinement, at least in the primary visual cortex (Ko et al. 2013), which seems to coincide in time with the reduction in SWA. Therefore, an intriguing possibility would be that cortical synaptic refinement, rather than synaptic pruning, may play a role in the developmental decline in SWA.

To investigate if and how synaptic refinement can be sufficient to explain observed changes in SWA during development, we resorted to computer simulations in which we could systematically manipulate synaptic refinement while assessing the resulting effects on SWA. The simulations were performed by taking advantage of a large-scale model of the primary visual cortex and associated thalamic regions that reproduces accurately cortical activity during wake and sleep (Esser et al. 2005, 2007; Hill and Tononi 2005; Olcese et al. 2010). In the present work, neural activity in the model was assessed in its wake and sleep mode both before and after implementing synaptic refinement in a way that mimicked physiological data (Ko et al. 2013). We show that, when we rewired lateral corticocortical connections such that neurons with similar POs and RFs were linked more selectively (hardwired refinement), there was invariably a marked decline in SWA in accord with experimental data in different species. We further show that an equivalent decline of SWA could be obtained by achieving synaptic refinement through spike-timing dependent plasticity (STDP), where connections

underwent repeated cycles of potentiation through exposure to visual stimuli in the wake-mode, followed by synaptic renormalization.

Materials and Methods

The large-scale model used in this study was first introduced by Hill and Tononi (2005). The following describes the organization, connectivity, and cellular and synaptic parameters of the version used herein. Previous model versions have been used to: (i) produce realistic wake and sleep dynamics (Hill and Tononi 2005); (ii) reproduce realistic responses to both visual stimuli (Hill and Tononi 2005) and transcranial magnetic stimulations (Esser et al. 2005); (iii) predict the breakdown of effective connectivity during sleep (Esser et al. 2009); (iv) assess the effects of changes in synaptic strength on slow wave amplitude and slope (Esser et al. 2007); and (v) observe the effects of homeostatic synaptic renormalization on learning and memory (Olcese et al. 2010). The present model version closely resembles that of Esser et al. (2007) in parameters and connectivity. However, herein only V1 is simulated, with some adaptations so that local excitatory connections could undergo synaptic refinement similar to that observed *in vivo*. A list of all network connections and model parameters is in Tables 1-3.

Table 1. Connectivity

Source Layer	Cell Type	Target Layer	Cell Type	Transmitter	Style	Pmax	Radius/size	Strength	Mean Delay	St. Dev Delay
Optic nerve	Exc	LGN	Exc, Inh	AMPA	Gaussian	0.75	1	10	2	1
Thalamic										
LGN	Exc	NRT	Exc	AMPA	Gaussian	1	2	2	2	0.25
LGN	Inh	LGN	Exc, Inh	gabaA	Gaussian	1	2	2	1	0.25
NRT	Inh	NRT	Inh	gabaA	Gaussian	0.5	12	2	1	0.25
NRT	Inh	LGN	Exc, Inh	gabaA	Gaussian	0.15	12	2	2	0.25
NRT	Inh	LGN	Exc, Inh	gabaB	Gaussian	0.05	12	2	2	0.25
M-cells	Exc	M-cells	Exc, Inh	AMPA, NMDA	Rectangular	0.03	10x10	1	2	0.25
M-cells	Inh	M-cells	Exc, Inh	gabaA	Gaussian	0.3	2	2.9	2	0.25
Thalamocortical										
LGN	Exc	IV, Infra	Inh	AMPA	Gaussian	0.1	5	3.5	3	0.25
LGN	Exc	IV, Infra	Exc	AMPA	Rectangular	0.5	8x1	3.5	3	0.25
M-cells	Exc	Supra, Infra	Exc,Inh	AMPA	Gaussian	0.1	12	0.3	7	0.2
Corticothalamic										
Infra	Exc	NRT	Inh	AMPA	Gaussian	0.5	5	0.5	8	0.5
Infra	Exc	LGN	Exc, Inh	AMPA	Gaussian	0.25	7	0.33	8	0.5
Infra	Exc	M-cells	Exc, Inh	AMPA	Gaussian	0.1	12	0.62	5	1
V1 intraareal										
Supra	Exc	Supra	Exc	AMPA	Rectangular	0.064	10x10	1	2	0.25
Supra	Exc	Supra	Exc	NMDA	Rectangular	0.064	10x10	0.75	2	0.25
Supra	Exc	Supra	Inh	AMPA	Gaussian	0.05	12	0.6	2	0.25
Supra	Inh	Supra	Exc, Inh	gabaA	Gaussian	0.025	7	0.5	2	0.25
Supra	Inh	Supra	Exc, Inh	gabaB	Gaussian	0.5	2	0.5	2	0.25
IV	Exc	IV	Exc	AMPA	Rectangular	0.0384	10x10	1	2	0.25
IV	Exc	IV	Inh	AMPA	Gaussian	0.05	12	0.6	2	0.25
IV	Inh	IV	Exc, Inh	gabaA	Gaussian	0.025	7	0.5	2	0.25
Infra	Exc	Infra	Exc	AMPA	Rectangular	0.0384	10x10	1	2	0.25
Infra	Exc	Infra	Inh	AMPA	Gaussian	0.05	12	0.6	2	0.25
Infra	Exc	Infra	Exc, Inh	gabaA	Gaussian	0.025	7	0.5	2	0.25
V1 interareal										
Supra	Exc	Infra	Exc,Inh	AMPA	Gaussian	1	2	1	2	0.25
Supra	Exc	Infra	Exc	NMDA	Gaussian	1	2	1	2	0.25
Supra	Inh	IV, Infra	Exc, Inh	gabaB	Gaussian	0.5	2	0.5	2	0.25
IV	Exc	Supra	Exc, Inh	AMPA	Gaussian	1	2	2.21	2	0.25
Infra	Exc	Supra	Exc, Inh	AMPA	Gaussian	1	2	2.21	2	0.25
Infra	Exc	IV	Exc, Inh	AMPA	Gaussian	1	2	2.21	2	0.25

Table 2. Neuron spike parameters

Neuron Type	θeq, mV	Tθ , ms	tspike, ms	Tspike, ms	Tm, ms
Cortical Exc	-51	2	2	1.75	16.0
Cortical Inh	-53	1	1	-0.5	8.0
Thalamic	-53	0.75	1	0.75	8.0

Table 3. Intrinsic neuronal currents

Channel	Neurons	Value-Waking	Value-Sleeping
g_{NaP}	Exc/Inh	0.5	1.25
g_h	Exc	0.5	4.0
g_T	30% Exc	0.5	4.0
g_{DK}	Exc/Inh	0.5	0.8
g_{KL}	Exc/Inh	1.0	1.85

Overall organization. The large-scale model (Fig. 3.1) consists of a section of V1 (14,400 simulated neurons) along with an associated section of thalamus (2,400 neurons). The model design enables realistic physiological responses, such as edge detection, when presented with simulated retinal stimuli embedded in noise (Hill and Tononi 2005; Esser et al. 2007).

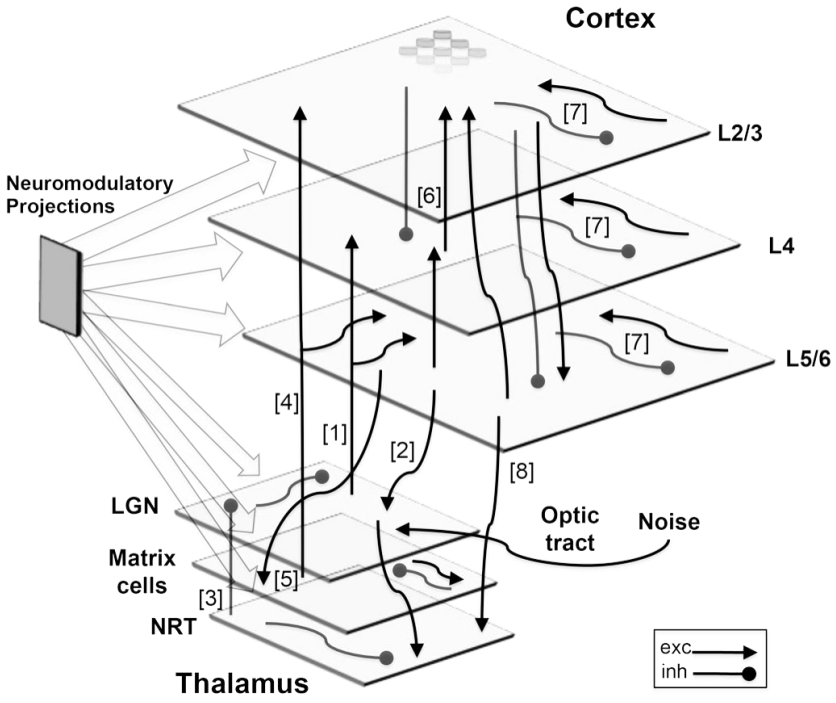


Figure 3.1. Schematic diagram. V1 and a thalamic area consisting of the LGN, matrix cells, and NRT. Visual inputs of either presented stimuli or background noise are projected onto the LGN via the optic tract (right). Thalamocortical loops are implemented by projections from the LGN to L4 and L5/6 [1], and excitatory back connections from L5/6 to the LGN [2]. The NRT sends inhibitory projections diffusely to neurons in the LGN [3]. Matrix cells provide broad excitation to both L2/3 and L5/6 layers [4], and receive diffuse excitatory feedback from L5/6 [5]. These loops, along with broad lateral excitatory connections between matrix cells, allow matrix cells to act as a global oscillator coupled to the cortex. Cortical interlaminar loops constructed of highly specific columnar projections [6] go from L4 to L2/3, from L2/3 to L5/6, and then from L5/6 back to L2/3 and L4, enacting vertical interlaminar processing. Cortical intralaminar lateral connections [7], both excitatory and inhibitory, span each layer. Feedback connections from the cortex [8] regulate the activity of the thalamus. There are numerous diffuse neuromodulatory connections (left). Network not drawn to scale.

The simulated cortex has 3 layers: the supragranular layer (L2/3), layer 4 (L4), and the infragranular layer (L5/6), each with its own inter and intra-laminar connectivity. Cortical columns are represented by groups of 9 model neurons (2 excitatory and 1 inhibitory neuron per layer) (Mountcastle 1997). The simulated V1 area is retinotopically structured and corresponds to approximately 0.6 cm^2 of striate cortical surface (equating to a monocular patch of 6x6 degrees in the parafoveal visual field). This means that locations on the cortex have corresponding locations in the thalamus and retina, a relationship mediated by afferents from the model lateral geniculate nucleus (LGN). These afferents converge onto individual neurons in both L4 and L5/6 (creating 8x1 RFs), and enforce POs. Four different POs of bars are explicitly implemented in the model: 0° (vertical), 45° (right-slanting diagonal), 90° (horizontal), and 135° (left-slanting diagonal).

The thalamus is composed of 3 distinct cellular populations, each with its respective function: core cells in the LGN, matrix cells, and intrathalamic inhibitory cells in the reticular thalamic nucleus (NRT). Each topographic location in the LGN is composed of 2 neurons: 1 that acts as an X-relay cell and the other as an inhibitory interneuron. For simplicity, only the ON portion of the RFs of the thalamic neurons is modeled. Matrix cells project diffusely to cortex (Jones and Hendry 1989). Hence, model matrix cells diffusely target L2/3 and L5/6, receive feedback via excitatory corticothalamic connections from L5/6, and assist global synchronization during sleep. The NRT also receives corticothalamic feedback from L5/6, as well as

excitatory inputs from the LGN, while at the same time providing the LGN with inhibitory feedback.

Connectivity. Presynaptic neurons connect probabilistically to postsynaptic neurons. Connection profiles are either Gaussian spatial density profiles (as for local inhibitory connections), or rectangular, uniformly distributed profiles (as for the thalamocortical afferents that implement POs and specified RFs). There are two categories: focused (area < 8x8, for example interlaminar connections) or diffuse (area ≥ 8x8, for example lateral inhibitory connections). Although the ratio of excitatory to inhibitory neurons in the model is 2:1, to resemble *in vivo* data, the ratio of excitatory to inhibitory synapses is 80/20 (White and Keller 1989). Conduction delays are taken into account in the connection profiles. Full details of connections profiles are in Table 1.

Model Neurons. The model is composed of single-compartment spiking neurons implementing Hodgkin-Huxley-style currents. The simplified dynamics of the fast spiking currents (I_{Na} and I_K) preserve the efficiency of integrate-and-fire neurons for large-scale simulations (Hill and Tononi 2005). The change in subthreshold membrane potential *V* for each neuron is given by:

$$\frac{dV}{dt} = [-g_{NaL}(V - E_{Na}) - g_{KL}(V - E_K) - I_{syn} - I_{int}]/\tau_m - g_{spike}(V - E_K)/\tau_{spike}$$

The sodium leak conductance (g_{NaL} = 0.4; E_{Na} = 30 mV) and potassium leak conductance (g_{KL} = 1.0; E_K = -90 mV) are the primary contributors to the resting membrane potential. All neural parameters are from Hill and Tononi (2005), including the activation of a fast potassium current during a spike in the form of a brief pulse (in which g_{spike} = 1 for a

duration t_{spike}), the governing time for the fast hyperpolarizing current (τ_{spike}), the threshold time constant (τ_θ) that determines the time to return to the equilibrium threshold, and the membrane time constants (τ_m), all summarized in Table 2. The membrane potential is influenced by synaptic input (I_{syn}) and intrinsic currents (I_{int}). A spike occurs when the membrane potential *V* reaches threshold, which drives *V* to the sodium reversal potential E_{Na}. After a spike, neurons enter a refractory period of length τ_θ. Rapid synaptic depression occurs through depletion of presynaptic vesicle pools (Zucker and Regehr 2002; Hill and Tononi 2005).

The synaptic input, I_{syn}, is a summation of all incoming currents from synaptic channels:

$$I_{syn} = \sum_{i,j} g_j^{(i)}(t)(V - E_j)$$

The amplitude and time course of each current is defined by the time-varying conductance *g(t)*, for each afferent *i*, on each channel *j*. The reversal potential E_j for each channel determines its sign. Each current has a rise and decay time constant (τ_1 and τ_2). Excitatory current comes from voltage-independent (AMPA) and voltage-dependent (NMDA) channels, while inhibition occurs via fast ($GABA_A$) and slow ($GABA_B$) channels. Specifically: AMPA (τ_1 = 0.5 ms, τ_2 = 2.4 ms, E_{rev} = 0 mV), NMDA (τ_1 = 4.0 ms, τ_2 = 40.0 ms, E_{rev} = 0 mV) $GABA_A$ (τ_1 = 1.0 ms, τ_2 = 7.0 ms, E_{rev} = -70 mV in the cortex and -80 mV in thalamic cells), and $GABA_B$ (E_{rev} = -90 mV). The synaptic parameters of $GABA_B$ are fully explicated in Esser et al. (2009), and are designed to undergo nonlinear changes in conductance in response to consistent activation, based on a previous channel model (Destexhe and Sejnowski 1995). Except for the updated $GABA_B$ channels, all parameters are as described in Hill and Tononi (2005). However,

the g_{peak} values of the model neurons were adapted to sustain single excitatory postsynaptic potentials of ~1 mV amplitude and inhibitory postsynaptic potentials of 1-1.5 mV amplitude (values in Table 3).

Intrinsic currents, I_{int}, also influence the firing dynamics of all cortical and thalamic neurons (as described in Hill and Tononi (2005)). The conductances of these intrinsic currents are listed in Table 3. Briefly, the currents included in the model are a) a non-inactivating hyperpolarization-activated cation current I_h, which underlies a depolarizing "pacemaker" potential in thalamic and cortical cells (Huguenard and McCormick 1992; McCormick and Bal 1997), b) a low-threshold fast-activating calcium current I_T, involved in burst firing in the thalamus (Huguenard and McCormick 1992), c) the persistent sodium current I_{NaP}, which activates at a sub-threshold voltage and inactivates on the order of seconds, d) a depolarization-activated potassium current I_{DK}, which acts as a generalized Na^+ or Ca^{2+}-activated K^+ current, and is known to play an important role in the UP and DOWN states of slow waves (Sanchez-Vives and McCormick 2000; Steriade 2003), and e) the potassium leak current I_{KL}, which is critical for the transition between wake and sleep modes of firing (Hill and Tononi 2005).

Transitioning from wake to sleep is triggered by the simulated change in neuromodulator levels, which act on cholinergic, noradrenergic, serotonergic, histaminergic, and glutamatergic metabotropic receptors in cortex and thalamus (McCormick 1992). The corresponding changes in model parameters are given in Table 3. The primary driver of the wake to sleep transition is the increase in potassium leak conductance (from 1.0 to 1.85), which simulates the effect of reduced release of arousal neuromodulators like acetylcholine and norepinephrine during sleep.

Sources of Spontaneous Activity and Visual Stimuli. Background noise in the model originates from spontaneous optic tract firing. This firing is modeled as 400 separate Poisson processes, and is independent of the wake/sleep cycle. The optic nerve cells project onto the LGN. From there the background noise percolates throughout the network, producing irregular spontaneous activity in all layers of the model (see Fig. 3.2). Another source of noise are "minis," spontaneously released neurotransmitter quanta (Vautrin and Barker 2003).

Visual stimuli were presented over retinothalamic projections onto the LGN with a strength (w) of 50 for a duration of 100 ms. This caused strong responsive firing of a topographic section of the thalamus the same size and shape as the presented bar.

Synaptic Refinement. Synaptic refinement is implemented in the thalamocortical model by modifying intralaminar, lateral connections. In a first set of simulations, connections in all cortical layers were rewired in a predefined manner to reflect experimental data (Ko et al. 2013). In the pre-refinement model, the lateral intralaminar excitatory connections among cortical neurons are distributed uniformly, irrespective of PO and only diffusely connected over RFs. Comparatively, in the post-refinement model neurons connect preferentially to neurons with similar POs and RFs (details in Results). Inhibitory connections were not altered during refinement, since inhibitory cells in the adult cortex show significantly less orientation selectivity in their lateral connections than excitatory cells do (Kisvarday 1997).

In a second set of simulations, excitatory intralaminar cortical connections in all cortical layers were augmented with a version of spike-timing dependent plasticity (STDP). In order to generate refinement via activity-dependent plasticity, visual stimuli were presented, bringing about a shift in synaptic strength profiles (see Results). The STDP rules were taken from Olcese et al. (2010), where they were used to examine the homeostatic regulation of synaptic strength. Long-term potentiation is thought to involve the AMPA receptor (Collingridge et al. 2004), along with the NMDA receptor as a triggering mechanism (Brader et al. 2007). Therefore STDP is modeled to affect the synaptic strength *w* of postsynaptic AMPA and NMDA receptors, which in turn modulates the synaptic peak conductances (g_{peak}) of excitatory connections (Abbott and Nelson 2000; Dan and Poo 2004). The particular STDP weight-change equation is adopted from Standage et al. (2007):

$$\Delta w_{p,d} = k_{p,d} m_{p,d}(w) e^{-c_{p,d}\Delta t}$$

where Δw is the change in synaptic weight, *p* and *d* stand for potentiation and depression, respectively, *k* is the learning rate, Δt is the time difference between the post and pre-synaptic spikes, and *m* is a weight-dependent factor, which ensures that small synapses will have greater percentage changes than large ones. In particular,

$$m_{p,d} = a_{p,d} w^{b_{p,d}}$$

with $a_p = 431$, $a_d = -59$, $b_p = 0.4$, $b_d = 0.1$, $c_p = 0.039$, $c_d = 0.043$, $k_p = 6e^{-5}$, $k_d = -6e^{-5}$ (Standage et al. 2007).

Plastic changes can depend on the intracellular calcium level (entering via NMDARs (Brader et al. 2007). Thus, as in Olcese et al. (2010), no plastic changes occur for low levels of calcium currents; at the level typical of spontaneous activity, the

standard STDP rule applies and there can be both potentiation and depression, while at a higher Ca^{2+} concentration only potentiation occurs.

Reduced model of isolated supragranular cortex section. To examine a range of connectivity changes and their effect on SWA, we simulated sleep activity in a smaller-scale cortical model: 2,700 neurons from supragranular cortex of the large-scale model in isolation. Strengths of individual connections were adjusted to ensure similar wake and sleep behavior as in the large-scale model (Table 4). Inputs from the rest of the cortex were substituted with noise in the form of 900 separate Poisson processes. Initially, every neuron connected to every other neuron with equal probability. Starting from this situation of complete homogeneity, we gradually increased connection selectivity and observed the resulting effect on SWA (see Results for details).

Table 4. Small-scale model connectivity

Source Layer	Cell Type	Target Layer	Cell Type	Transmitter	Style	Pmax (height)	Radius or size	Strength	Mean Delay	St. Dev Delay
Cortical noise	Exc	Cortex	Exc, Inh	AMPA	Gaussian	0.75	1	7.5	2	1
Cortex										
Cortical sheet	Exc	Cortical sheet	Exc, Inh	AMPA	Rectangular	0.01	30x30	1	2	0.25
Cortical sheet	Exc	Cortical sheet	Exc, Inh	NMDA	Rectangular	0.01	30x30	0.35	2	0.25
Cortical sheet	Inh	Cortical sheet	Exc, Inh	gabaA	Rectangular	0.01	30x30	1.5	2	0.25
Cortical sheet	Inh	Cortical sheet	Exc, Inh	gabaB	Rectangular	0.01	30x30	0.5	2	0.25

Data Gathering and Analysis. In all the simulations models were started with identical initial states. The membrane potentials of cortical and thalamic neurons, as well as the synaptic currents of cortical neurons, were recorded at every simulation time step (4000 Hz). Recording began after 10s of the sleep mode to let the simulation settle to its equilibrium behavior.

Since the cortical EEG is believed to reflect postsynaptic currents in cortical pyramidal cells (Buzsáki et al. 2012), an approximate EEG of the model's V1 was

derived by summing individual postsynaptic currents over all V1 model neurons. To analyze SWA and detect individual slow waves, the EEG was band-pass filtered (at 0.5 - 30 Hz, stopband edge, frequencies 0.5 - 80 Hz, stopband minimal attenuation 10 dB) using a Chebyshev Type II filter design, as in Esser et al. (2007). The total simulation of 20s was broken into non-overlapping epochs of 4-second length. To detect individual slow waves the EEG signal of each epoch was centered around zero by subtracting the mean value. Slow waves were defined as positive (> 0) deflections between 2 consecutive negative (< 0) peaks. Wave amplitude was measured as the maxima within the positive deflection. Waves with amplitudes below 20% of the global peak amplitude for that epoch were excluded to prevent small fluctuations around the zero-line counting as slow waves. The slope of a wave was defined as the average of the slope of the first wave segment (from the preceding negative peak to the positive peak) and second-wave segment (from the positive peak to the subsequent negative peak).

Simulation techniques. Simulations were run using *Synthesis*, an object-orientated neural simulator (https://github.com/Caanon/Synthesis). Using a quad-core Mac Pro running Mac OS 10.4 with 9 GB RAM, each second of simulation took around 30 minutes to compute. The Runge-Kutta 4^{th} order method was used to perform numerical integration over a step size of 0.25 ms. Running at a lower step size did not affect previous results (Esser et al. 2009). Herein, changes in initial conditions and small percentile changes to synaptic parameters and connectivity numbers did not alter the main experimental findings. The software package MATLAB (The Math Works, Inc., Natick, MA) was used for analysis.

Results

In the following simulations we show that synaptic refinement impacted both the waking and sleeping behavior of the thalamocortical model, resulting in a drop in SWA and changes in slow wave morphology. Synaptic refinement affected the lateral intralaminar excitatory connections of all 3 cortical layers in the model, and was accomplished by rewiring in a predefined manner. Next, in a further set of simulations, we investigated how synaptic plasticity can bring about synaptic refinement by changing profiles of synaptic strength. Finally, a smaller model is used to make a general case that the greater amount of synaptic refinement there is, the larger the drop in SWA.

Sleep/wake activity prior to synaptic refinement. The pre-refinement thalamocortical model (Fig. 3.1; described in Materials and Methods) corresponds to a pre-adolescent period early on in brain development. During the wake mode the model engages in low-voltage spontaneous activity, which resembles physiological waking activity in the EEG (Fig. 3.2*A*, top). Cellular membrane potentials fluctuate rapidly around -60 mV, and rarely hyperpolarize below -70 mV. Spontaneous activity differs in frequency and intensity across cortical layers. Deep layers fire more frequently during wake, while superficial layers show sparser firing (Fig. 3.2*B*), as observed in the real brain (Niell and Stryker 2010; Sakata and Harris 2012).

In the pre-refinement model, the POs and RFs of cortical neurons are already in place due to selective thalamocortical feedforward connections. Local cortical connections, however, are still aspecific (Espinosa and Stryker 2012). Topographic plots

of post-stimuli neuronal spikes demonstrate that the RFs and POs of simulated cortical neurons are indeed intact and functional prior to refinement (3.2*C*). Since V1 is retinotopic, when a bar is presented to the optic nerve, the population response of neurons with a matching PO reflects the shape of the bar. The presentation of a 0° bar, for example, generated activity in the population of L4 neurons with a PO of 0° (Fig. 3.2*C*). Neurons not selective for the stimulus do not show any evoked response above baseline spontaneous activity.

The same retinothalamic afferents that carry information about visual stimuli also originate spontaneous activity. For this reason, POs and RFs were reflected in the spontaneous cortical firing. In Fig. 3.2*B*, for example, snapshots of the membrane potential of populations in L4 and L5/6 with a preference for 0° (vertical) bars are shown, and exhibit subthreshold fluctuations in the form of vertical bars. This result is in line with experimental data from the visual cortex in non-human primates showing that cellular membrane potentials of a neuronal population during spontaneous activity reflect the functional properties of that population (Tsodyks 1999; Kenet et al. 2003).

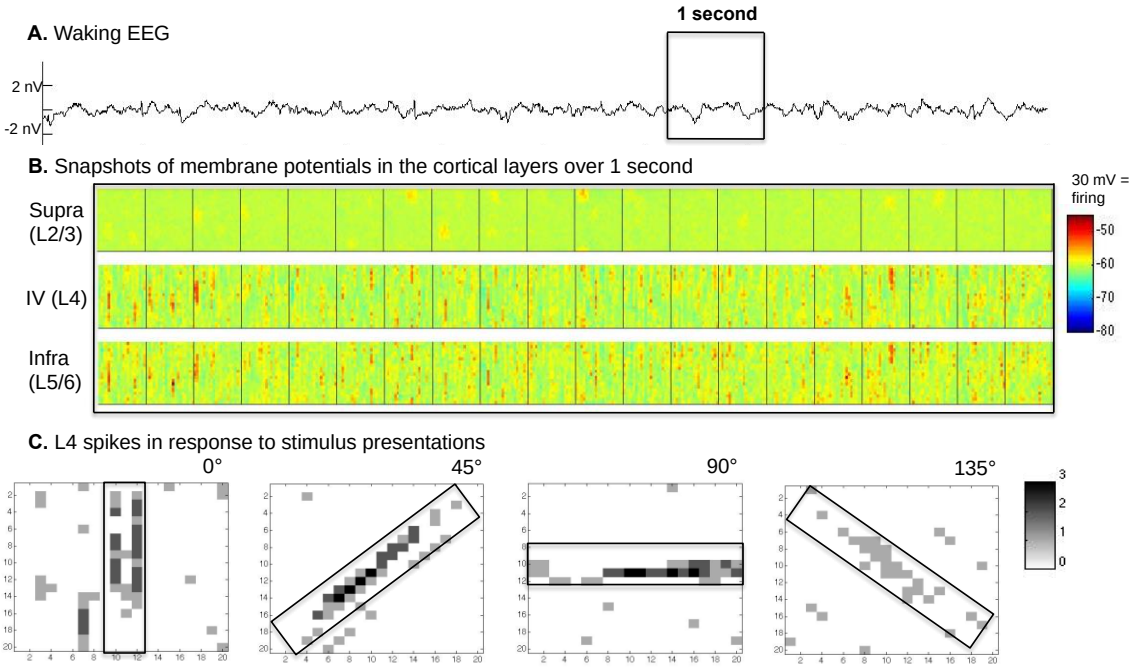


Figure 3.2. Simulated waking activity prior to refinement: spontaneous and evoked. Synaptic currents and membrane potentials for all neurons were recorded during a 10s simulation under the waking mode. *A*: EEG without stimulus presentation. *B*: membrane potential snapshots of a 20x20 neuron cortical patch over the boxed 1s, with an image captured every 50 ms from excitatory populations in L2/3, L4, and L5/6. *C:* The matrix plots show the summed spiking response of 4 20x20 populations of excitatory L4 neurons during the 100 ms after their preferred stimulus was presented through the optic tract (size and shape of the bar presented are indicated by the rectangle over the analogous retinotopic area on the cortex).

In the sleep mode, the pre-refinement model exhibits high-amplitude SWA. Fig. 3.3*A* shows the persistent generation of oscillations in the simulated EEG, a prototypical slow wave pattern with clearly delineated global UP and DOWN states. The EEGs from populations of neurons with different POs are synchronous in their entrance and exit of UP and DOWN states (Fig 3.3*B*). The entire thalamocortical system participates in these slow waves, assisted by matrix cells, which act in the thalamus to help widely excite cortical neurons into the UP state (Fig. 3.3*C*). The DOWN states are characterized by network silence (no or little firing) and globally hyperpolarized membrane potentials (~ -

80 mV). UP states are characterized by globally depolarized membrane potentials (~ -58 mV) and elevated firing throughout the network. These 2 states can also be observed in the snapshots of the membrane potentials (taken every 50 ms for 1s of sleep) across the different populations with distinct POs (Fig. 3.3*D*). This behavior is characteristic of the bistability of neuronal activity during deep sleep (Steriade et al. 2001; Tononi and Massimini 2008).

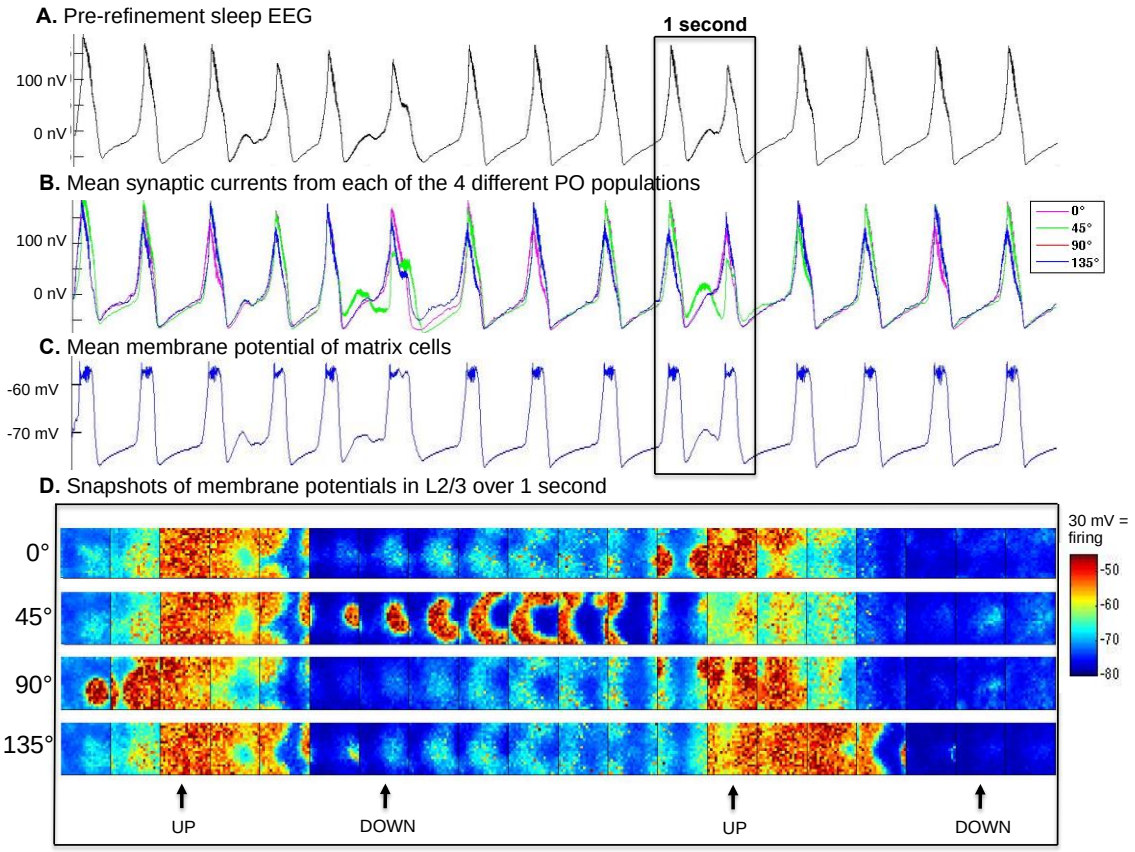


Figure 3.3. Simulated activity during sleep prior to refinement. Synaptic currents and membrane potentials for all neurons were recorded during a 20s simulation in the sleep mode. *A*: 10s of EEG. *B*: Mean synaptic current of the populations in the supragranular layer with different POs, plotted individually. *C*: Mean membrane potential of thalamic matrix cells. *D*: Snapshots of the supragranular membrane potential of 20x20 populations over 1s, with an image captured every 50 ms from the 4 populations with different POs. Snapshots show the origination and termination of UP states (with an average membrane potential of around -60 mV). A neuron fires when its membrane potential crosses its dynamic threshold, after which its membrane potential is set to +30 mV. The neurons that fired in each snapshot can thus be identified as the darkest red dots. Note that only a

small subset of neurons is firing at any given time during the UP state.

Reorganization of connections simulates synaptic refinement. In order to reflect a post-adolescent period in brain development, the intralaminar connections of the model were refined. This refinement, which was done by rewiring the connections, was based on physiological data from the mouse visual cortex (Ko et al. 2013). Accordingly, the post-refinement model was created from the pre-refinement model by manipulating the model's probabilistic connectivity profiles, keeping all other parameters identical. In the pre-refined model, which has highly homogenous lateral connectivity, neurons connect with equal probability over a local cortical area (10x10) and are insensitive to their targets' POs (25% to each of the 4 POs on average). Refinement consisted of rewiring these intralaminar connections, so that the probability of a neuron being connected to the same PO doubled from 25% to 50%, while connection probabilities to neurons with 45° and 90° differences decreased from 25% to 19% and 11%, respectively. Crucially, the overall net number of connections remained unchanged (Fig. 3.4*A*). In addition, lateral connections were refined from a diffuse to a focused connectivity profile, by decreasing the profile area by ~50% (10x10 to 7x7). Since similar RFs are spatially proximal to each other, this leads to a post-refinement increase in the likelihood of connections between neurons with more similar RFs (Fig. 3.4*A*). These wiring changes were applied to excitatory neurons in all cortical layers.

Ko et al. (2013) assessed refinement by studying supragranular neurons in the visual cortex of mice at different stages of development. Critically, those same neurons were then identified in slices so as to assess local connectivity. This was done by stimulating target neurons and finding evoked EPSPs in their neighbors (2 to 6 proximal

neurons within 50 μm). The local connectivity of neurons became more selective for neurons with similar RFs and POs over development. To demonstrate that the changes in connectivity profiles in the model are congruent with these experimental observations, a similar experiment was conducted: a single target neuron was stimulated to fire for 40 ms, while the membrane potentials of 6 neighboring cells were recorded (Fig. 3.4*B*). Two of the 6 neighboring neurons showed rising responses during stimulation. Since in the model the connectivity matrices of all connection types are available, it could be verified that these 2 neighbors (in blue) indeed received lateral excitatory connections from the target cell, while the other 4 (in black) had no connections. This demonstrates that, as *in vivo*, the model's individual excitatory connections can be identified via single-cell stimulations. To further verify that the model's true connection probabilities can, in principle, be inferred from local recordings (as done in Ko et al. (2013)) lateral excitatory connections of 140 randomly selected supragranular neurons were examined with respect to 6 nearby neurons (Fig. 3.4*C*). The mean probability of finding a connection (post-refinement: mean = 0.10, SEM = 0.017; pre-refinement: mean = 0.07, SEM = 0.017) was not significantly different before and after refinement ($P = 0.08$, 2-sample t-test). Prior to refinement, lateral excitatory connectivity showed no significant preference for similar POs ([$F = 0.92$, $P = 0.4$], ANOVA). Post-refinement, neurons were found to connect preferentially to their own orientation and aversely to their opposite orientation ([$F = 0.92$, $P < 0.0001$], ANOVA). The locally assessed connection probabilities for both the pre- and post-refinement models overlap with the actual imposed model values used in the model's construction (marked as red 'x's in Fig. 3.4*C*, right). The total connection probabilities are similar to those observed experimentally (Ko et al. 2013), although

connectivity is overall sparser in the model. Our results show that locally derived connection probabilities indeed approximate the true values.

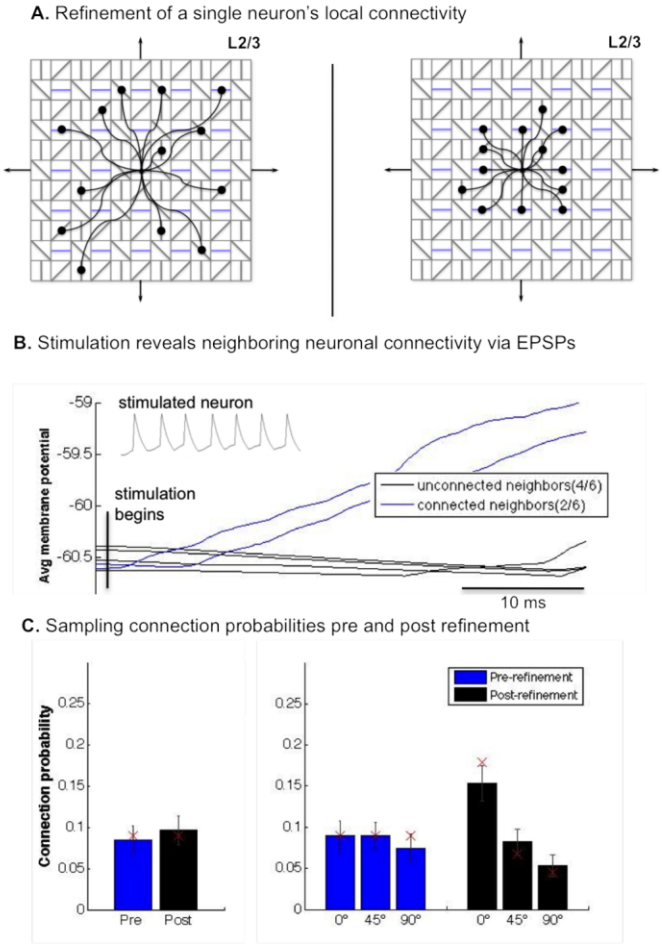


Figure 3.4. Refinement of neural connectivity. *A*: A topographic schematic of the refinement of an example L2/3 neuron (center) with a 90° PO. Prior to refinement (left) the neuron's 13 local excitatory connections were probabilistically distributed uniformly across all 4 distinct POs (25% each) over a 10x10 area around the projecting neuron. Post-refinement (right), it has doubled its connections to neurons that share its PO (in blue), while decreasing its connections to those with a different PO. Additionally, the area of local connectivity has been reduced by ~50% to 7x7, increasing the probability of it connecting to neurons with similar RFs. *B*: Mean responses of 6 neurons adjacent to a neuron that was repeatedly stimulated (so as to ensure an action potential) for 40 ms across 15 stimulation trials. Recordings aligned to stimulus onset. *C*: Mean connection probabilities of 140 randomly selected supragranular neurons with respect to 6 nearby neurons (connections assessed directly). Error bars denote SEM. Red 'x's denote the absolute probabilities of lateral excitatory connections.

Responsiveness to stimuli in the post-refinement model. While spontaneous waking activity in the post-refinement model is similar to the pre-refinement model (low-amplitude and variable), we examined whether refinement changed how the model responds to stimuli. Both the pre- and post-refinement model were presented with a series of 15 vertical (0°) bars over their retinothalamic afferents. This reliably resulted in selective responses in L4 (after propagation through the LGN). The total number of spikes within the appropriate retinotopic zone of L4 was recorded for 100 ms after each bar presentation (post-refinement: mean evoked spikes = 151.5, std = 57.57; pre-refinement: mean evoked spikes = 142, std = 60.47). To assess whether information transmission changed due to refinement, the signal-to-noise ratio (SNR) of the response was defined as:

$$SNR = \frac{mean(evoked\ spikes)}{std(spontaneous\ spikes)}.$$

Only spikes of neurons with the appropriate PO and within the appropriate retinotopic area were included. The appropriate retinotopic areas were defined as a 3x20 patch of L4 in the population with the PO that matched the presented bar. This patch corresponded topographically to where the bar was presented (as in Fig. 3.2*C*). The SNR for the pre-refinement condition was 14.76, while the post-refinement SNR increased to 17.97. By focusing lateral connectivity and increasing the connectivity between more similar POs, synaptic refinement may affect neuronal evoked responses, increasing the SNR of signal transmission.

Synaptic refinement leads to a reduction in SWA. The post-refinement model shows attenuated slow waves in the cortical EEG (Fig. 3.5*A*). The reason is that, in the post-

refinement model, supragranular populations with different POs enter UP and DOWN states less synchronously (Fig. 3.5*B*). However, global slow waves could still be observed due to periods of greater and lesser synchrony, as seen in the EEG and the response of the thalamic matrix cells coupled to the cortex (Fig. 3.5*C*). Yet, depolarization (~ -58 mV) during the UP states and hyperpolarization (~ -80 mV) during the DOWN states is no longer global, as can be seen by membrane potential rasters (Fig. 3.5*D*).

SWA was measured over 20s of simulation, broken into 5 epochs (see Materials and Methods). SWA significantly decreased in the post-refinement model (Fig. 3.5*E*); a reduction to 27% of the original pre-refinement SWA value (post-refinement SWA mean = 0.47, SEM = 0.28; pre-refinement SWA mean = 1.74, SEM = 0.01; $P < 0.01$, 2-sample t-test). As can be seen in Fig. 3.5*F*, this decrease in power spans most of the lower spectrum.

There is evidence that the morphology of individual slow waves reflects the cortical dynamics underlying the EEG. Decreased amplitude and slope signify a decreased ability of the slow wave to recruit neurons and spread along the cortex (Esser et al. 2007). Indeed, the developmental decrease in SWA is accompanied by a decrease in amplitude and slope in both humans (Riedner et al. 2007) and rats (Vyazovskiy et al. 2007).

To analyze these waveform properties, individual slow waves in the simulated EEGs were identified (see Materials and Methods). Slow waves in both the pre and post-refinement conditions (post-refinement $n = 55$; pre-refinement $n = 27$) were evaluated in terms of amplitude and slope. Refinement decreased the average wave amplitude to 33% of the pre-refinement value (Fig. 3.6*G*; post-refinement peak amplitude mean = 50.72 nV,

SEM = 4.45 nV; pre-refinement peak amplitude mean = 157.24 nV, SEM = 4.39 nV; $P < 0.01$, 2-sample t-test). The slope of the individual waves also decreased significantly after refinement (post-refinement: mean slope = 1.03 μV /s, SEM = .026 μV /s; pre-refinement: mean slope = 1.57 μV /s, SEM = 0.053 μV /s; $P < 0.01$, 2-sample t-test).

Since experimental data on synaptic refinement was obtained from the supragranular layer only (Ko et al. (2013)), it is not yet clear to what extent refinement occurs in all cortical layers. For this reason, we also created a post-refinement model wherein solely the L2/3 lateral excitatory connections were rewired in the manner described. As before, the result was a drop in SWA during the sleep mode, however this drop was attenuated compared to when all the layers were refined (just L2/3 refinement: post-refinement SWA mean = 0.61, SEM = 0.02; pre-refinement SWA mean = 1.74, SEM = 0.01; $P < 0.01$, 2-sample t-test). Similarly, refining just L2/3 reduced wave amplitude (post-refinement peak amplitude mean = 69.71 nV, SEM = 3.78 nV; pre-refinement peak amplitude mean = 157.24 nV, SEM = 4.39 nV; $P < 0.01$, 2-sample t-test) and slope (post-refinement: mean slope = 0.9 μV /s, SEM = .057 μV /s; pre-refinement: mean slope = 1.57 μV /s, SEM = 0.053 μV /s; $P < 0.01$, 2-sample t-test), but to a lesser degree than refining all 3 cortical layers.

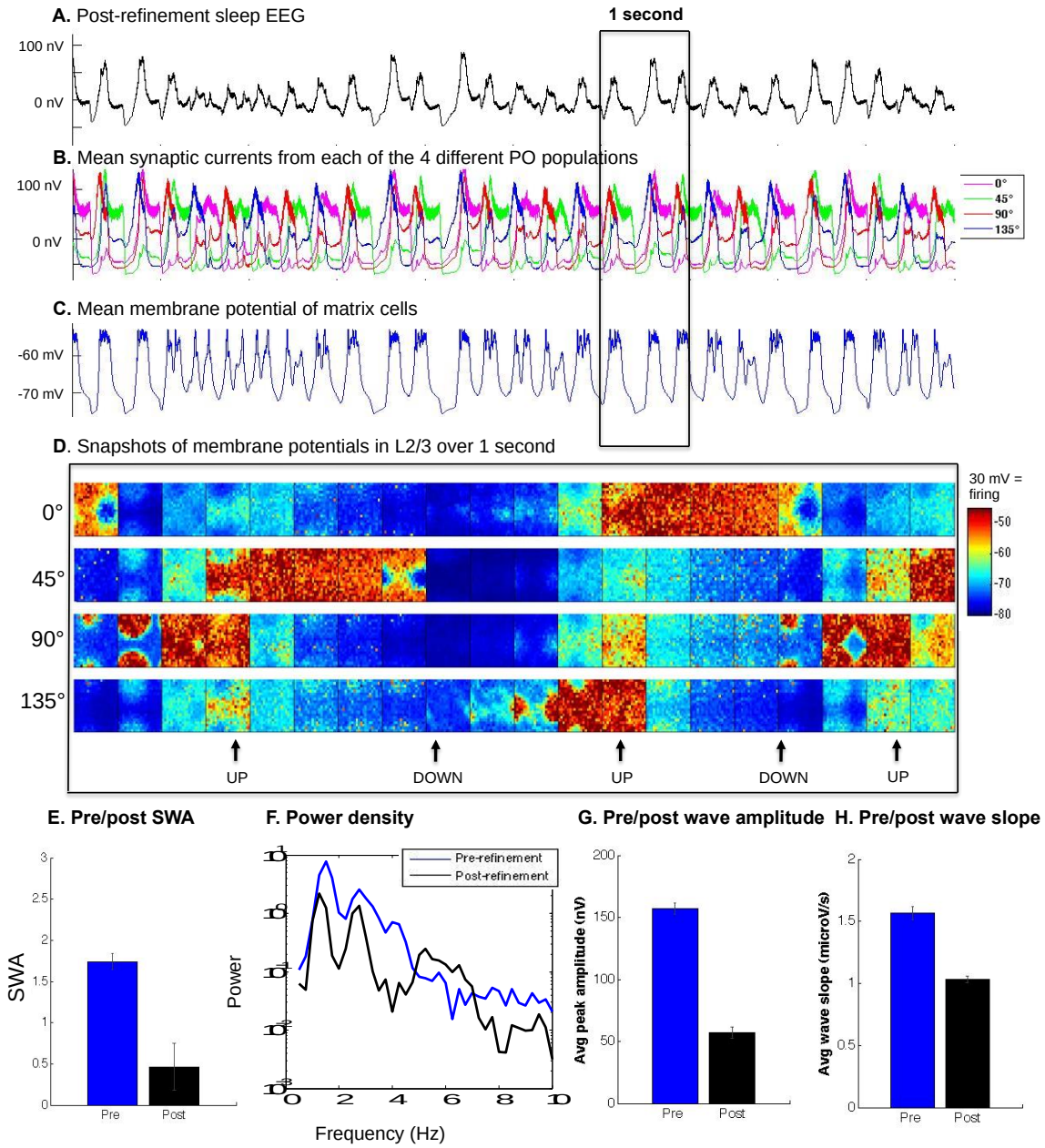


Figure 3.5. Simulated activity during sleep post-refinement. *A*: 10s of EEG, showing attenuated slow waves. *B*: Mean synaptic currents of the populations in the supragranular layer with different POs, plotted individually. Populations are much less synchronous in their activity, as well as more variable in the timing of their UP states, than in the pre-refinement model (compare to Fig. 3.3*B*). *C*: Mean membrane potential of the thalamic matrix cells. *D*: Snapshots of the supragranular membrane potential of 20x20 populations over 1s, with an image captured every 50 ms from the 4 populations with different POs. *E*: Post-refinement model shows a decrease in SWA. *F*: The lower-frequency power spectrum of the pre-refinement and post-refinement conditions averaged across sleep

epochs. *G-H*: Amplitude and the slope have decreased post-refinement. Error bars denote SEM.

Modeling the developmental process of synaptic refinement via STDP. It has been shown that an increase in the selectivity of connections during development can occur via synaptic plasticity, as changes in synaptic strength are forerunners for changes in connectivity (Sanes and Yamagata 2009). And indeed, there is evidence that synaptic refinement involves a rearrangement in the strengths of connections, as neurons in adult mouse V1 with similar RFs and POs have stronger excitatory synaptic connections (Cossell et al. 2015). To explore this aspect of synaptic refinement, a second set of simulations tested whether the implementation of STDP in the lateral excitatory connections of the pre-refinement model could lead, after a period of learning, to synaptic refinement, and whether this in turn would result in a drop in SWA (see Materials and Methods for details).

During normal development, learning takes place over many cycles of wake followed by sleep. According to the synaptic homeostasis hypothesis (SHY, Tononi and Cirelli 2003, 2014), wake globally increases synaptic strength, and then during sleep there is synaptic renormalization in the form of a global reduction of synaptic strength back to a baseline level. This renormalization is best performed during sleep, when the brain is offline and spontaneous activity can sample the brain's extant knowledge in a statistically unbiased manner. Thus, learning according to SHY involves both local potentiation and global renormalization across a repeating wake/sleep cycle.

To mimic this process, we used the previously defined pre-refinement model (now referred to as pre-learning) and trained it over 11 wake sessions. In each training session the model was presented with 70 3x20 bar stimuli of all 4 POs at different

topographical locations. Each training session led to net potentiation. After each one of the 11 sessions we renormalized the average total synaptic weight back to its starting value pre-session, mimicking the effect of a period of sleep. This renormalization was done by decreasing the strength *w* of all neurons in a population with a shared PO proportionally to that population's increase in strength *w* during the prior training session. Renormalization was implemented only over the same lateral connections augmented with STDP. Critically, this meant that at the end of the 11 sessions the total synaptic strength of these connections was unchanged.

The results of these 11 sessions was synaptic refinement in the form of a rearrangement of synaptic strength without a change in the total amount. Neurons in all layers began, pre-learning, with a mean homogenous connection strength profile: [25%, 25%, 25%, 25%] to neurons with POs of 0°, 45°, 90°, 135° degrees difference to their own PO. Post-learning in the supragranular layer, this distribution had been shifted to a mean of [46%, 16.4%, 19.5%, 16.3%,]. Other layers showed near-identical results. Thus learning via STDP resulted in neurons connecting most strongly to other neurons with shared PO (0° difference), as seen in Fig. 3.6*A*.

In the previous set of simulations (Figs 3.4, 3.5), synaptic refinement was accomplished by directly rewiring the model's connections. To compare the two cases: in the rewired model the mean connection profile shifted from [25%, 25%, 25%, 25%] pre-refinement to [50%, 19%, 11%, 19%] in the post-refinement model. If an increase of 1% in a synapse's strength corresponds to an increase of 1% in the chance that another connection is added, the refinement in the form of a rearranged synaptic strength is very similar to the refinement done by rewiring (as in Fig 3.4). STDP also potentiated local

connections between topographically, and thus functionally, closer neurons. Synaptic renormalization also led to a renormalization of connections between topographically, and thus functionally, distant neurons (Fig. 3.6*B*). The result was, again, similar to the focusing of local excitatory connectivity when synaptic refinement was accomplished by rewiring (Fig. 3.4*A*).

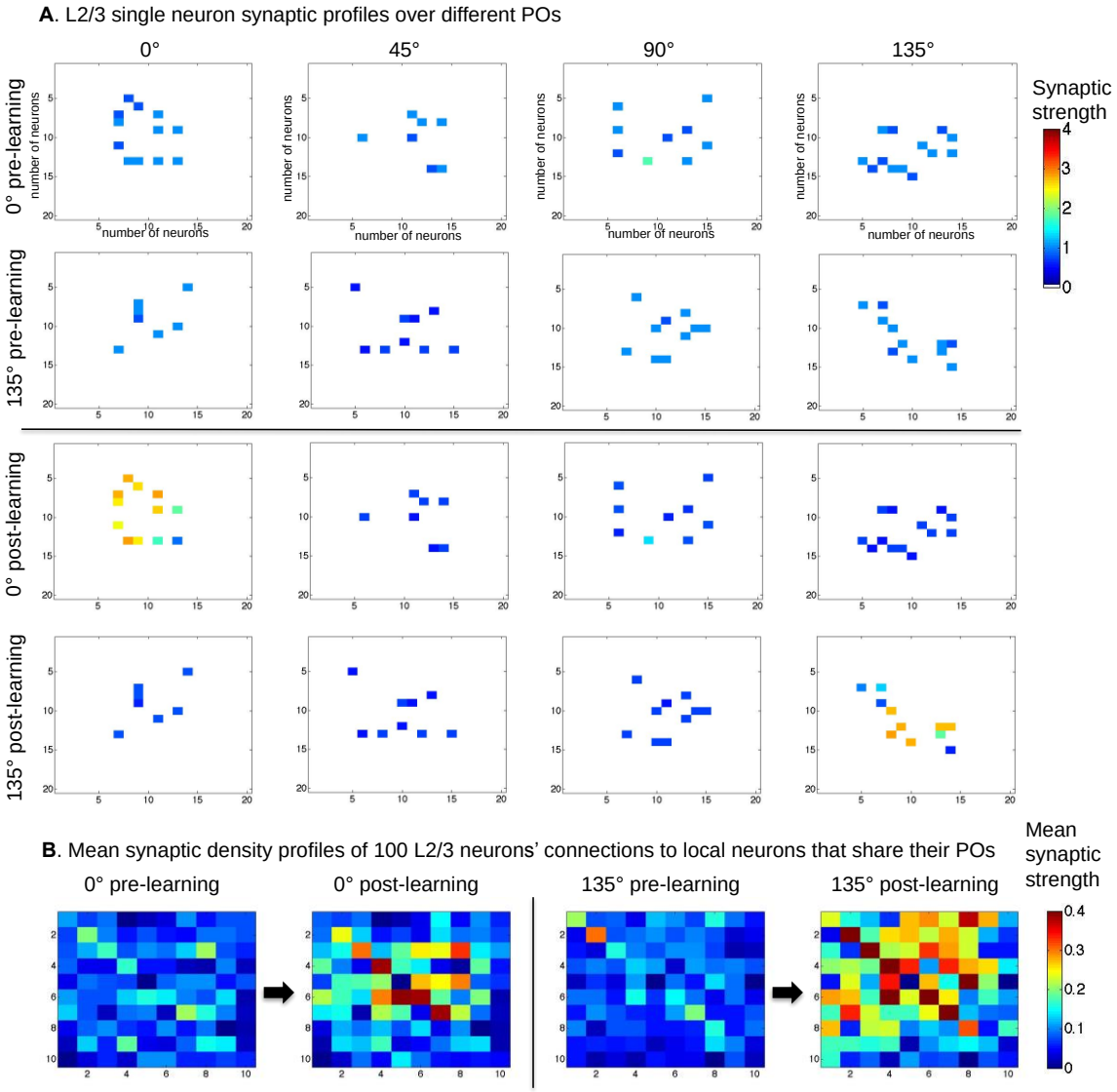


Figure 3.6. Learning-induced changes in synaptic strength. *A*: Synaptic strength profiles of the projecting connections of 2 individual L2/3 neurons with POs of 0° and 135°, respectively. Shown in this figure are the changes to the local connections of 2 individual neurons, one with a 0° preferred orientation and the other with 135° to neighboring neurons of all the different preferred orientations, pre- and post-learning. Each row shows

the neuron's projections and their synaptic strength over its 20x20 nearest neighbors of each particular preferred orientation. Pre-learning (top), the strengths of both neurons' projecting connections were uniform across POs (a mean of str = 1, std = 0.25). Post-learning (bottom), the 2 neurons connect preferentially to neurons that share their POs. Total synaptic strength remains the same. *B*: Mean synaptic density profiles of the local excitatory connectivity of 100 neurons with a PO of 0° and another 100 with a PO of 135°. Sampling was over a 10x10 area centered on each neuron, and was conducted both pre and post learning. This sampling reveals the focusing of lateral connectivity via learning.

The effects of learning on SWA. After synaptic refinement was brought about via learning across wake/sleep cycles, STDP was disabled and the post-learning model was run in the sleep mode for 20s while all synaptic currents and membrane potentials were recorded. Just as in the post-refinement model, slow waves in the post-learning model decreased in amplitude (Fig. 3.7*A*-*H*). Performing the same analysis over 5 4s epochs of the sleep mode, the post-learning model also showed a drastic decrease in the amount of SWA (post-learning SWA mean = 0.093, SEM = 0.008; pre-learning SWA mean = 1.74, SEM = 0.01; $P < 0.01$, 2-sample t-test). Resultant slow waves (post-learning $n = 69$; pre-learning $n = 27$) were analyzed in terms of amplitude and slope, which both showed significant reductions (post-learning peak amplitude mean = 33.44 nV, SEM = 1.47 nV; pre-learning peak amplitude mean = 157.24 nV, SEM = 4.39 nV; $P < 0.01$, 2-sample t-test; post-learning: mean slope = 1.13 μV /s, SEM = 0.08 μV /s; pre-learning: mean slope = 1.57 μV /s, SEM = 0.053 μV /s; $P < 0.01$, 2-sample t-test). These results suggest that a set of cycles of potentiation followed by renormalization can lead to synaptic refinement, and this can have a profound effect on SWA.

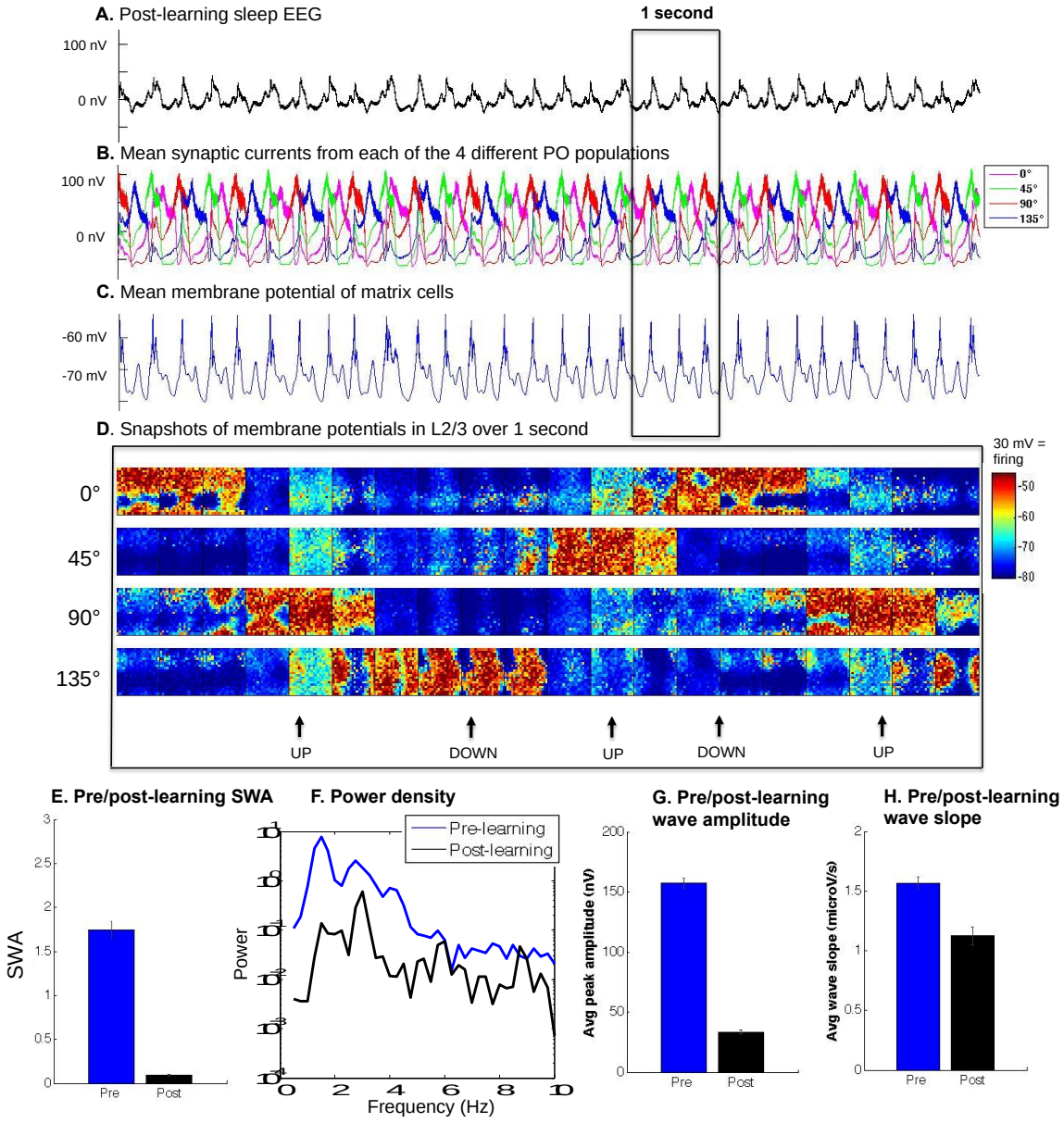


Figure 3.7. Refinement via synaptic plasticity reduces SWA. *A*: The EEG of 10s of slow wave sleep post-learning. *B*: Similar to the post-refinement condition, post-learning the mean synaptic currents from populations with different POs are more asynchronous. *C*: Mean membrane potential of thalamic matrix cells. *D*: Snapshots of the supragranular membrane potential of a 20x20 population over 1 second, with an image captured every 50 ms from populations with POs of 0°, 45°, 90°, and 135°, respectively. Note the similarity in dynamics to Fig 3.5*D*. *E*: SWA drops post-learning. *F*: The lower-frequency power spectrum of the pre-learning and post-learning conditions averaged across sleep epochs. *G-H*: Both the amplitude and the slope have decreased in the post-learning condition. Error bars denote SEM.

Examining the effects of topographic and orientation refinement in a model section of the supragranular layer. In the large-scale thalamocortical model it would be computationally unfeasible to systematically explore how changes in the size and type of refinement impact SWA. To that end, a 2,700-neuron section of the supragranular layer of the large-scale model was excised and modified so that it displayed normal wake/sleep behavior (see Materials and Methods; Table 4 for details). As in the large-scale model, only lateral excitatory connections were rewired during refinement.

Synaptic refinement, as modeled here, has two aspects: a focusing of lateral connectivity, which increases topographic selectivity (Callaway and Katz 1991) and an increase in orientation selectivity of connections (Ko et al. 2013). Initial connectivity of the L2/3 cortical section was homogenous: each neuron connected diffusely to the entire 30x30 model area (100% network coverage), and no neuron had a preference for any PO.

To isolate the effect of focusing lateral connectivity on SWA, the area that neurons connected to in a probabilistic manner was reduced from 30x30 to 15x15, 7x7, 5x5, 4x4, and 3x3. However, the net number of connections remained constant across all conditions. The average membrane potential (used in lieu of the EEG as the signals are nearly identical) was recorded in each condition of varying connectivity over a 10s period of sleep. The first 2s of activity were excluded from analysis to allow the simulation to settle. The focusing of lateral connections caused SWA to drop off sharply once the spread of the lateral connections was smaller than 5x5 (Fig. 3.8*A*).

The second aspect of synaptic refinement modeled herein is the increase in orientation selectivity. Selectivity is when populations of neurons have a greater amount of connectivity among themselves than with the rest of the brain; across entire brain

regions this is typically referred to as modularity (Meunier et al. 2009). Such modules can be groups of neurons that share an orientation, but also any commonly activated or functionally similar groups. To formalize this, we here define the selectivity (S) of a network with discrete populations as a ratio between two connectivity values: the difference between the number of actual connections established by groups of neurons and the number of connections that those neurons would establish if the network were randomly wired, compared to the maximum possible difference. Complete homogeneity, or random connectivity, thus corresponds to S = 0, while at S = 1, neurons only have connections to other neurons in their population.

To investigate the effect of increasing orientation selectivity, neurons in the L2/3 section were randomly assigned to different populations with different POs (representing different modules). Then the S of the network was increased in stages from S = 0 to S = 0.5, 0.9, 0.95, 0.975, and 1, for conditions of 2, 4, and 8 distinct populations with different POs. Again the total number of neuronal connections stayed the same in all cases, but here the distance of the lateral connections also stayed the same. Increasing S decreased SWA (Fig. 3.8*B*), and did so more effectively the higher the number of distinct POs. With only 2 POs, SWA remained basically unchanged up to S = 0.95. When the section had 8 distinct POs SWA decreased substantially from S = 0.5 on.

To test the interaction between these two aspects of synaptic refinement, we reduced the lateral connectivity of the section to the area it was in the post-refinement thalamocortical model (7x7, which here is 5% network coverage by each neuron). The network was given 4 POs and run through increases in S. Combining the two aspects of refinement hastened the reduction in SWA (Fig. 3.8*C*).

The decreases in SWA were the result of asynchrony between neural populations. Fig. 3.8*D* shows the average membrane potentials of the 4 populations with different POs in the S = 0.9, 7x7 lateral connectivity condition (marked by a red 'x' in Fig. 3.8*C*). This condition is, in the dynamics of its slow waves, closest to the large-scale post-refinement model. As can be directly seen, slow waves decreased in amplitude due to averaging effects.

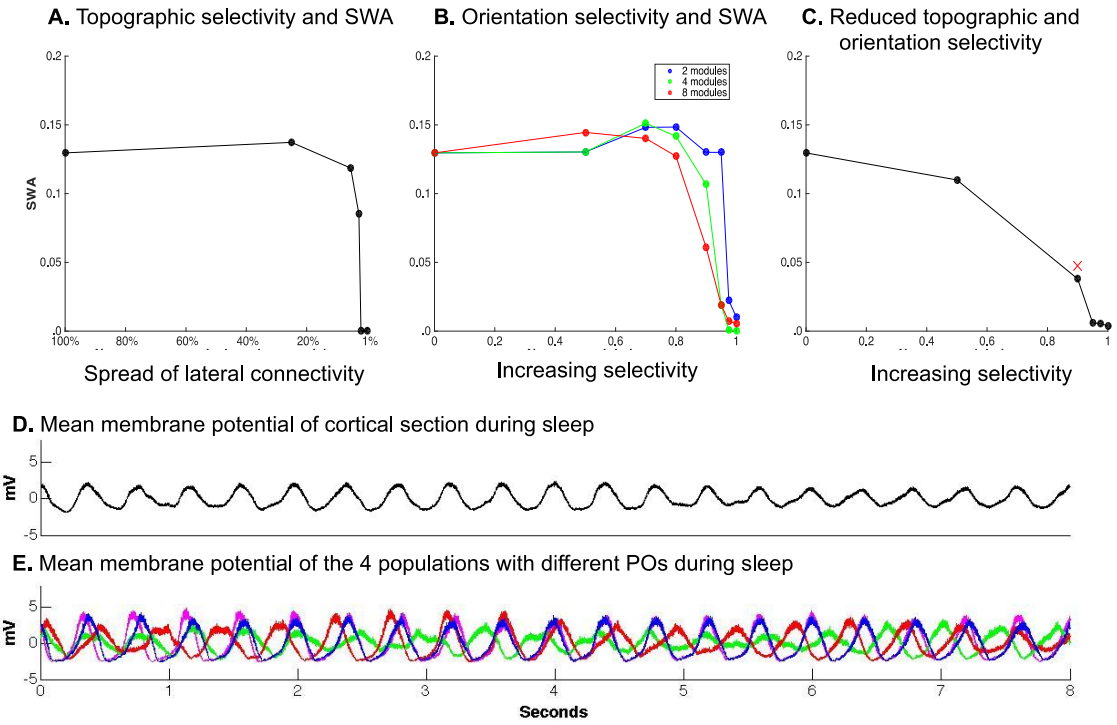


Figure 3.8. Refining by degrees. *A*: The effects of decreasing the spread of lateral connectivity on SWA. *B*: The effects of increasing the selectivity between populations with different POs. *C*: With lateral connectivity fixed at 5% coverage, a 4 PO network was run through conditions of increasing orientation selectivity. In this combined condition, the drop-off in SWA appeared earlier and is more pronounced than in (*A*) and (*B*). *D*: Mean membrane potential of the entire network over 8s of sleep for the condition marked as a red 'x' in *C*. *E*: In the same condition, mean membrane potentials of the 4 populations with distinct POs. Note the asynchrony of UP and DOWN states.

Note that the increase in selectivity required for a reduction in SWA comparable to experimental data (~50%) is higher (0.9) in the small-scale model than in the large-

scale model (de Vivo et al. 2014; Jenni and Carskadon 2004). In the large-scale model, refinement was implemented as a doubling of connections to the same preferred orientation, consistent with Ko et al. (2013), corresponding to a value of increased selectivity of ~0.33. This difference is not surprising given the smaller size and the lack of thalamocortical and corticocortical feedback of the reduced model. Nevertheless, the small-scale model allows to make a general point, that the greater the selectivity, the greater the reduction in SWA.

Discussion

We show here that synaptic refinement as implemented in our simulations is sufficient to cause a marked decrease in SWA. Specifically, simulating synaptic refinement by rewiring the model's connections directly (Fig. 4) led to a 27% reduction in SWA (Fig. 5). This decline is close in magnitude to the 50% reduction in SWA observed in the frontal and parietal cortices in adolescent mice (de Vivo et al. 2014), as well as to the 47% drop in SWA observed in EEG recordings during adolescent development in humans (Jenni and Carskadon 2004). Additionally, the decline in amplitude and slope of slow waves was also similar in magnitude to that seen in humans (Kurth et al. 2010a). Of note, refinement in the model led to a drop in EEG power that was not restricted to SWA but extended to all frequencies below 10 Hz, a finding also consistent with results *in vivo* in both rodents (Olini et al. 2013; de Vivo et al. 2014) and humans (Buchmann et al. 2011).

Experimental data indicate that synaptic refinement requires visual experience, and that it is the corticocortical connectivity that is refined post eye-opening, not the

feedforward thalamocortical connectivity (Ko et al. 2013). For example, in dark-reared rats, cortical networks do not fully develop the appropriate corticocortical selective connectivity (Ishikawa et al. 2014). To explore how visual experience might bring about synaptic refinement, we implemented STDP in the local excitatory connections of the model. Molecular, anatomical, and physiological studies have shown that synaptic plasticity during waking experience leads to a net increase in synaptic strength, whereas periods of sleep lead to synaptic renormalization (Tononi and Cirelli 2014). To mimic this process, we exposed the model to stimuli in a "wake" mode, leading to net synaptic potentiation. Exposure to stimuli was then followed by synaptic renormalization. This cycle was repeated over multiple sessions. The results of the simulations suggest that a series of wake/sleep cycles can bring about the selective reorganization of synaptic strength that accompanies synaptic refinement (Cossell et al. 2015), which is in turn associated with a drop in SWA (Figs 6, 7). Previous work has indicated that such repeated patterns of synaptic potentiation during wake and then renormalization during sleep can also have beneficial effects on memory at both the system level (Nere et al. 2013) and at the level of individual neurons (Hashmi et al. 2013).

Overall, the present results support the hypothesis that synaptic refinement may contribute significantly to the drop in SWA observed during adolescence, and outline a mechanism of action in the form of daily cycles of potentiation/renormalization. An important remaining question is whether the developmental decrease in SWA parallels the time courses of synaptic refinement. Evidence from mice shows decreases in SWA from at least P20 onwards (de Vivo et al. 2014), a period of time overlapping with the occurrence of synaptic refinement (P13 – P26; Ko et al. 2013). In our simulations SWA

tracks refinement non-linearly, beginning to decrease only once a critical amount of refinement has occurred (Fig. 8). Thus, refinement may be ongoing throughout brain development but only become relevant to SWA after it reaches a critical level.

Across cortex, refinement during development is most likely not restricted to groups of neurons with similar orientation selectivity and receptive fields, but will affect many functional groups of neurons. The fact that a greater number of different groups of neurons facilitates the drop in SWA (Fig. 8*B*) indicates that in cortical areas that are very heterogeneous or contain many modules, smaller increases in selectivity may be required to account for the drop in SWA seen *in vivo*.

The model was designed to be as physiologically realistic as possible (see Methods), and our results are compatible with the evidence from intracellular recordings performed in cats, which showed that neuronal activity is less rhythmic and synchronous across cortical regions in sleep relative to anesthesia (Chauvette et al. 2011). The general results of this study also proved insensitive to small manipulations of the experimental parameters. Yet, the model has limitations. Thus, the duration of the intracellular DOWN states of neurons in the model (~150-200 ms) are the same as those observed during slow wave sleep *in vivo* (Chauvette et al. 2011), but the durations of the UP states (150-250 ms) are shorter. Moreover, the declines in SWA after refinement or after sleep tend to be larger than those seen in vivo. One reason for these discrepancies is likely that the parameters were necessarily taken from different species, and while the model was originally interpreted as a small portion of cat V1, the empirical parameters to emulate synaptic refinement were based on mice, and the rate of slow waves in the EEG is close to that found in rodents (Vyazovskiy et al. 2007). Moreover, we are modeling only a

single region, V1, which limits slow waves in their spread. Nevertheless, few other models are capable of producing physiologically realistic SWA together with appropriate visual responses in wake, as well as STDP. For example, the small model used in Ko et al. (2013) to examine synaptic refinement did not generate slow waves. Additionally, the model's robustness was tested by systematically varying the relevant parameters and initial conditions (amounting to over 3 years CPU time) when it was first introduced (Hill and Tononi 2005). The same model, under similar conditions to those herein, has previously been used for studying SWA (Hill and Tononi 2005; Esser et al. 2007; Olcese et al. 2010). One of these studies showed that the amount of SWA tracks global synaptic strength in the model (Esser et al. 2007). We build on these results by showing that changes in the local distribution of synaptic strength, from homogeneous to heterogeneous, can have similar effects. These effects may combine, or one or the other may dominate at different times, to drive developmental changes in SWA.

Of course, computer simulations cannot by themselves prove that synaptic refinement is the ultimate cause of the developmental decline in SWA, and it remains possible that other developmental processes such as synaptic pruning play a major role. However, the present results are consistent with the idea that it is possible to indirectly track the maturation of cortical connectivity during development by monitoring the progressive decline of SWA – a relatively simple, non-invasive procedure that can be performed at multiple time points (Ringli and Huber 2011; Ringli et al. 2013; Cirelli and Tononi 2015). For example, it is known that in humans the developmental decline in SWA begins in posterior brain regions and moves to anterior regions (Kurth et al. 2010b). Based on the present results, one would predict that synaptic refinement may progress in

a similar posterior-to-anterior manner. Moreover, one would predict that alterations in SWA maturation may reflect changes in synaptic refinement that may underlie some developmental disorders.

Acknowledgements

The authors thank Umberto Olcese, Sean Hill, and Tim Blakely for answering questions about Synthesis.

Grants

Funded by NIMH grant R01MH099231 and NINDS grant P01NS083514 to GT and CC. LA and EH were funded by the Templeton World Charities Foundation (Grant #TWCF 0067/AB41).

Disclosures

G. Tononi consults for Philips Respironics and is involved in a research study in humans supported by Philips Respironics. This study is not related to the work presented in the current manuscript. The other authors have indicated no financial conflicts of interest.

References

Abbott LF, Nelson SB. Synaptic plasticity: taming the beast. *Nat Neurosci* 3 Suppl: 1178–1183, 2000.

Blakemore S-J, Choudhury S. Development of the adolescent brain: implications for executive function and social cognition. *J Child Psychol Psychiatry* 47: 296–312.

Brader JM, Senn W, Fusi S. Learning real-world stimuli in a neural network with spike-driven synaptic dynamics. *Neural Comput* 19: 2881–912, 2007.

Buchmann A, Ringli M, Kurth S, Schaerer M, Geiger A, Jenni OG, Huber R. EEG sleep slow-wave activity as a mirror of cortical maturation. *Cereb Cortex* 21: 607–15, 2011.

Buzsáki G, Anastassiou CA, Koch C. The origin of extracellular fields and currents--EEG, ECoG, LFP and spikes. *Nat Rev Neurosci* 13: 407–20, 2012.

Callaway EM, Katz LC. Effects of binocular deprivation on the development of clustered horizontal connections in cat striate cortex. *Proc Natl Acad Sci U S A* 88: 745–9, 1991.

Campbell IG, Feinberg I. Longitudinal trajectories of non-rapid eye movement delta and theta EEG as indicators of adolescent brain maturation. *Proc Natl Acad Sci U S A* 106: 5177–80, 2009.

Casali AG, Gosseries O, Rosanova M, Boly M, Sarasso S, Casali KR, Casarotto S, Bruno M-A, Laureys S, Tononi G, Massimini M. A theoretically based index of consciousness independent of sensory processing and behavior. *Sci Transl Med* 5: 198ra105, 2013.

Cirelli C, Tononi G. Cortical development, electroencephalogram rhythms, and the sleep/wake cycle. *Biol Psychiatry* 77: 1071–8, 2015.

Collingridge GL, Isaac JTR, Wang YT. Receptor trafficking and synaptic plasticity. *Nat Rev Neurosci* 5: 952–62, 2004.

Cossell L, Iacaruso MF, Muir DR, Houlton R, Sader EN, Ko H, Hofer SB, Mrsic-Flogel TD. Functional organization of excitatory synaptic strength in primary visual cortex. *Nature* 518: 399–403, 2015.

Craik FIM, Bialystok E. Cognition through the lifespan: mechanisms of change. *Trends Cogn Sci* 10: 131–8, 2006.

Dan Y, Poo M-M. Spike timing-dependent plasticity of neural circuits. *Neuron* 44: 23–30, 2004.

Destexhe A, Sejnowski TJ. G protein activation kinetics and spillover of gamma-aminobutyric acid may account for differences between inhibitory responses in the hippocampus and thalamus. *Proc Natl Acad Sci U S A* 92: 9515–9, 1995.

Espinosa JS, Stryker MP. Development and plasticity of the primary visual cortex. *Neuron* 75: 230–49, 2012.

Esser SK, Hill S, Tononi G. Breakdown of effective connectivity during slow wave sleep: investigating the mechanism underlying a cortical gate using large-scale modeling. *J Neurophysiol* 102: 2096–111, 2009.

Esser SK, Hill SL, Tononi G. Modeling the effects of transcranial magnetic stimulation on cortical circuits. *J Neurophysiol* 94: 622–39, 2005.

Esser SK, Hill SL, Tononi G. Sleep homeostasis and cortical synchronization: I. Modeling the effects of synaptic strength on sleep slow waves. *Sleep* 30: 1617–30, 2007.

Feinberg I. Schizophrenia: caused by a fault in programmed synaptic elimination during adolescence? *J Psychiatr Res* 17: 319–34, 1983.

De Felipe J, Marco P, Fairén A, Jones EG. Inhibitory synaptogenesis in mouse somatosensory cortex. *Cereb Cortex* 7: 619–34, 1997.

Gilbert C, Wiesel T. Clustered intrinsic connections in cat visual cortex. *J Neurosci* 3: 1116–1133, 1983.

Hashmi A, Nere A, Tononi G. Sleep-Dependent Synaptic Down-Selection (II): Single-Neuron Level Benefits for Matching, Selectivity, and Specificity. *Front Neurol* 4: 148, 2013.

Hill S, Tononi G. Modeling sleep and wakefulness in the thalamocortical system. *J Neurophysiol* 93: 1671–98, 2005.

Hoel EP, Albantakis L, Tononi G. Quantifying causal emergence shows that macro can beat micro. *Proc Natl Acad Sci U S A* 110: 19790–5, 2013.

Huguenard JR, McCormick DA. Simulation of the currents involved in rhythmic oscillations in thalamic relay neurons. *J Neurophysiol* 68: 1373–83, 1992.

Innocenti GM, Price DJ. Exuberance in the development of cortical networks. *Nat Rev Neurosci* 6: 955–65, 2005.

Ishikawa a. W, Komatsu Y, Yoshimura Y. Experience-Dependent Emergence of Fine-Scale Networks in Visual Cortex. *J Neurosci* 34: 12576–12586, 2014.

Jenni OG, Carskadon MA. Spectral analysis of the sleep electroencephalogram during adolescence. *Sleep* 27: 774–83, 2004.

Jones EG, Hendry SHC. Differential Calcium Binding Protein Immunoreactivity Distinguishes Classes of Relay Neurons in Monkey Thalamic Nuclei. *Eur J Neurosci* 1: 222–246, 1989.

Kandel ER, Markram H, Matthews PM, Yuste R, Koch C. Neuroscience thinks big (and collaboratively). *Nat Rev Neurosci* 14: 659–64, 2013.

Kenet T, Bibitchkov D, Tsodyks M, Grinvald A, Arieli A. Spontaneously emerging cortical representations of visual attributes. *Nature* 425: 954–6, 2003.

Kisvarday Z. Orientation-specific relationship between populations of excitatory and inhibitory lateral connections in the visual cortex of the cat. *Cereb Cortex* 7: 605–618, 1997.

Ko H, Cossell L, Baragli C, Antolik J, Clopath C, Hofer SB, Mrsic-Flogel TD. The emergence of functional microcircuits in visual cortex. *Nature* 496: 96–100, 2013.

Ko H, Hofer SB, Pichler B, Buchanan K a, Sjöström PJ, Mrsic-Flogel TD. Functional specificity of local synaptic connections in neocortical networks. *Nature* 473: 87–91, 2011.

Ko H, Mrsic-Flogel TD, Hofer SB. Emergence of feature-specific connectivity in cortical microcircuits in the absence of visual experience. *J Neurosci* 34: 9812–6, 2014.

Kurth S, Jenni OG, Riedner BA, Tononi G, Carskadon MA, Huber R. Characteristics of sleep slow waves in children and adolescents. *Sleep* 33: 475–80, 2010a.

Kurth S, Ringli M, Geiger A, LeBourgeois M, Jenni OG, Huber R. Mapping of cortical activity in the first two decades of life: a high-density sleep electroencephalogram study. *J Neurosci* 30: 13211–9, 2010b.

Lenroot RK, Giedd JN. Brain development in children and adolescents: insights from anatomical magnetic resonance imaging. *Neurosci Biobehav Rev* 30: 718–29, 2006.

Mayer-Schönberger V, Cukier K. *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt, 2013.

McCormick DA, Bal T. Sleep and arousal: thalamocortical mechanisms. *Annu Rev Neurosci* 20: 185–215, 1997.

McCormick DA. Neurotransmitter actions in the thalamus and cerebral cortex and their role in neuromodulation of thalamocortical activity. *Prog Neurobiol* 39: 337–88, 1992.

Meunier D, Achard S, Morcom A, Bullmore E. Age-related changes in modular organization of human brain functional networks. *Neuroimage* 44: 715–23, 2009.

Mountcastle VB. The columnar organization of the neocortex. *Brain* 120 (Pt 4: 701–22, 1997.

Nere A, Hashmi A, Cirelli C, Tononi G. Sleep-dependent synaptic down-selection (I): modeling the benefits of sleep on memory consolidation and integration. *Front Neurol* 4: 143, 2013.

Niell CM, Stryker MP. Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron* 65: 472–9, 2010.

Oizumi M, Albantakis L, Tononi G. From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Comput Biol* 10: e1003588, 2014.

Olcese U, Esser SK, Tononi G. Sleep and synaptic renormalization: a computational study. *J Neurophysiol* 104: 3476–93, 2010.

Olini N, Kurth S, Huber R. The effects of caffeine on sleep and maturational markers in the rat. *PLoS One* 8: e72539, 2013.

Paus T, Keshavan M, Giedd JN. Why do many psychiatric disorders emerge during adolescence? *Nat Rev Neurosci* 9: 947–57, 2008.

Petanjek Z, Judaš M, Šimic G, Rasin MR, Uylings HBM, Rakic P, Kostovic I. Extraordinary neoteny of synaptic spines in the human prefrontal cortex. *Proc Natl Acad Sci U S A* 108: 13281–6, 2011.

Riedner BA, Vyazovskiy V V, Huber R, Massimini M, Esser S, Murphy M, Tononi G. Sleep homeostasis and cortical synchronization: III. A high-density EEG study of sleep slow waves in humans. *Sleep* 30: 1643–57, 2007.

Ringli M, Huber R. Developmental aspects of sleep slow waves: linking sleep, brain maturation and behavior. *Prog Brain Res* 193: 63–82, 2011.

Ringli M, Souissi S, Kurth S, Brandeis D, Jenni OG, Huber R. Topography of sleep slow wave activity in children with attention-deficit/hyperactivity disorder. *Cortex* 49: 340–7, 2013.

Sakata S, Harris KD. Laminar-dependent effects of cortical state on auditory cortical spontaneous activity. *Front Neural Circuits* 6: 109, 2012.

Sanchez-Vives M V, McCormick DA. Cellular and network mechanisms of rhythmic recurrent activity in neocortex. *Nat Neurosci* 3: 1027–34, 2000.

Sanes JR, Yamagata M. Many paths to synaptic specificity. *Annu Rev Cell Dev Biol* 25: 161–95, 2009.

Spear LP. The adolescent brain and age-related behavioral manifestations. *Neurosci Biobehav Rev* 24: 417–63, 2000.

Sporns O, Chialvo DR, Kaiser M, Hilgetag CC. Organization, development and function of complex brain networks. *Trends Cogn Sci* 8: 418–425, 2004.

Standage D, Jalil S, Trappenberg T. Computational consequences of experimentally derived spike-time and weight dependent plasticity rules. *Biol Cybern* 96: 615–23, 2007.

Steriade M, Timofeev I, Grenier F. Natural Waking and Sleep States: A View From Inside Neocortical Neurons. *J Neurophysiol* 85: 1969–1985, 2001.

Steriade M. The corticothalamic system in sleep. *Front Biosci* 8: d878–99, 2003.

Tau GZ, Peterson BS. Normal development of brain circuits. *Neuropsychopharmacology* 35: 147–68, 2010.

Tononi G, Cirelli C. Sleep and synaptic homeostasis: a hypothesis. *Brain Res Bull* 62: 143–50, 2003.

Tononi G, Cirelli C. Sleep and the price of plasticity: from synaptic and cellular homeostasis to memory consolidation and integration. *Neuron* 81: 12–34, 2014.

Tononi G, Massimini M. Why does consciousness fade in early sleep? *Ann N Y Acad Sci* 1129: 330–4, 2008.

Tononi G. Consciousness as integrated information: a provisional manifesto. *Biol Bull* 215: 216–242, 2008.

Tononi G. Integrated Information Theory of Consciousness: An Updated Account. *Arch Ital Biol* 150: 56–90, 2012.

Tsodyks M. Linking Spontaneous Activity of Single Cortical Neurons and the Underlying Functional Architecture. *Science (80-)* 286: 1943–1946, 1999.

Vautrin J, Barker JL. Presynaptic quantal plasticity: Katz's original hypothesis revisited. *Synapse* 47: 184–99, 2003.

De Vivo L, Faraguna U, Nelson AB, Pfister-Genskow M, Klapperich ME, Tononi G, Cirelli C. Developmental patterns of sleep slow wave activity and synaptic density in adolescent mice. *Sleep* 37: 689–700, 700A–700B, 2014.

Vyazovskiy V V, Riedner BA, Cirelli C, Tononi G. Sleep homeostasis and cortical synchronization: II. A local field potential study of sleep slow waves in the rat. *Sleep* 30: 1631–42, 2007.

White E, Keller A. *Cortical circuits*. Boston, MA: Birkhauser, 1989.

White LE, Fitzpatrick D. Vision and cortical map development. *Neuron* 56: 327–38, 2007.

Yoshimura Y, Dantzker JLM, Callaway EM. Excitatory cortical neurons form fine-scale functional networks. *Nature* 433: 868–73, 2005.

Zucker RS, Regehr WG. Short-term synaptic plasticity. *Annu Rev Physiol* 64: 355–405, 2002.

Discussion

Overview of the three completed Aims

In summary, the work in Aim #1 demonstrated that, for any given system, it is possible to determine at which spatiotemporal grain causal power, measured as *effective information*, is maximal, and that this level is not necessarily the micro scale (causal emergence).

The work in Aim #2 extended this demonstration to a system's capacity to integrate information: an analysis that captures *irreducible* cause-effect power within the system. It showed that causal emergence could come about both via the macro having greater causal selectivity and also being more irreducible. The results open up the potential to empirically investigate at what spatiotemporal scale the brain's capacity for integrating information peaks (Tononi et al., in prep).

The work in Aim #3 used a large-scale thalamocortical model to investigate how brain organization affects SWA. We showed how a process of developmental synaptic refinement (implemented directly or through STDP in wake and synaptic renormalization in sleep) can account for the decrease in SWA that is observed during adolescence in multiples species. Now that this hypothesis is supported by our computer simulations, it may become possible for researchers to employ recordings of sleep SWA as an indicator for the occurrence or lack of occurrence of synaptic refinement in physiological and pathological conditions.

In all, these Aims set the stage for exciting projects that both continue the research already accomplished and also offer the possibility of combining the two lines of research (Aims #1 & 2 with Aim #3), by applying the developed measures of causal emergence to physiologically-realistic thalamocortical models.

Future direction: Black boxing

Aims #1 & 2 showed that causal emergence was possible, both in a generalized sense (using effective information), but also via the capacity of a system to integrate information. Key to the concept of causal emergence has been the notion of a macro element, defined along with its macro states. In the research that comprises this thesis a macro element was formalized as a group of micro elements, with a multiple-realizable macro state, formalized as a mapping over its underlying micro states (such that any information that can differentiate those micro states is lost at the macro level). That is, all macro states have so far been *coarse-grains* (Hoel et al., 2013; Hoel et al., in prep). However, coarse-graining may not be the only way to define a macro element's state. Rather, a recent research program in the lab has shown that a macro state can also be conceptualized similar in spirit to the approach proposed in cybernetics "black box" theory: parts (micro elements) of a macro element can be hidden from direct observation by black-boxing them (Ashby 1956), which sets the state of a macro element to that of its black-box output.

Black box theory runs contra to the widely held assumption in science that the more variables one can include in a causal model of a system, the better the model of that system is. For example, a "Big Data" approach is being used in areas as diverse as neuroscience (Kandel et al., 2013) and healthcare (Mayer-Schönberger and Cukier, 2013), and grows ever more attractive as computational power increases exponentially. Yet scientists often seem to purposefully leave out data in their descriptions and models. Consider, for example, that while molecular neuroscience has revealed that neurons have

highly complex internal dynamics, many research programs still view neurons as merely elements which take inputs (neurotransmitters) and which makes an internal decision and then produces an output at its axon (to spike or not to spike). That is, often neurons are implicitly considered as being black boxes (see Figure 4.1; taken from Marshall et al., in prep).

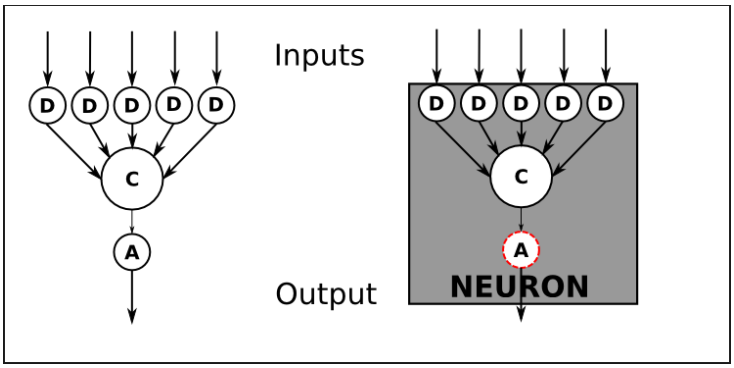


Figure 4.1: A causal model of a neuron constituted of micro elements (left) or taken as a black box (right). The neuron receives inputs at its dendrites (D), which are passed on to the cell body (C) and then axon (A), which outputs to other neurons. Alternatively, all these elements can be black boxed to form a single element that is the neuron.

In both approaches, black-boxing and coarse-graining, macro causal models have a reduced state-space compared to the micro causal model of a system. In black-boxing, the states of hidden micro elements are selectively ignored for the black box' macro state, while a coarse-grained macro state is a grouping and averaging of micro-level states. Nonetheless, each approach may increase a system's cause-effect power in their particular context. Neurons may be good candidates for being considered as black boxes. But if one looks at groups of neurons, particularly those with noisy and redundant activity and whose dynamics at a population level comprise reliable cause-effect relationships, such groups may be a good candidate for coarse-graining into even larger macros.

Preliminary results indicate that the capacity to integrate information can indeed peak when macro states are formed via black boxing. Black boxes appear to be able to beat their underlying micro level by revealing higher-order causal interactions between elements over multiple time-steps. The macro states created by coarse-graining, comparatively, as shown in Aims #1 & 2, beat their underlying micro level by being more deterministic and less degenerate in their causal interactions. Combining these two complimentary approaches paves the way for a unified definition of macro elements with macro states constructed by nested coarse-grains and black boxes. Together they will be more flexible and more powerful in their abilities to beat the underlying micro.

Future direction: Modeling and empirical approaches to spatiotemporal scale

Projects using the neural simulator Synthesis have consistently built on prior projects over the last decade (Hill and Tononi, 2005; Esser et al., 2007, 2009; Olcese et al., 2010). Next, Synthesis may be the neural simulator which provides the first *in silico* measures of integrated information in large-scale artificial neural systems. This is planned to take place via a grant (which was recently approved) led by Johan Storm, Steven Laureys, and Marcello Massimini, titled "Experimental and computational exploration of consciousness mechanisms and methods in mice and humans." Part of the grant will fund simulations that combine the data from the large-scale thalamocortical network of Synthesis with data from a biophysically detailed model of the rodent somatosensory cortex (developed by the Human Brain Project). The goal of these joint simulations is to use computational models to test ideas about basic principles and mechanisms for cortical integration and differentiation and, ultimately, for assessing consciousness.

Now is the time to pursue this critical project, since empirical heuristics for measuring integrated information have recently been developed, while its exact formulation is "NP-hard" and will thus remain computationally intractable for large neural networks (Oizumi et al., 2014). Comparisons to the simulated thalamocortical model are necessary to evaluate the accuracy of the proposed heuristics to the exhaustive measure of integrated information. For example, in human experiments, TMS can be used along with high-density electroencephalography (EEG) to estimate integrated information by way of the perturbational complexity index (PCI). PCI has been successful in estimating the level of consciousness of individual subjects across wakefulness, dreaming, NREM sleep, different levels of sedation via anesthetics, and in patients in different comatose stages (Casali et al., 2013). In previous work, it was shown that the large-scale thalamocortical model is capable of accurately reproducing neural responses to TMS (Esser et al., 2005). Hence, the stage is set for assessing the relation between PCI and other measures of integrated information by evaluating the model's responses to simulated TMS. Here a realistic thalamocortical model would allow investigating parameter manipulations that cannot be tested in vivo for practical or ethical reasons.

This research program is, at least in part, what was originally described as Aim #4 in the Thesis Proposal. Aim #4 planned to build upon Aims #1-3 to test the hypothesis that synaptic refinement during development is associated with an increase in the capacity for information integration. Its goal was to assess how the developmental changes in brain organization impact integrated information (possibly by changing at what spatiotemporal scale integrated information reaches a maximum). We hypothesized

that the process of synaptic refinement that occurs during neural development will be associated with an increase in integrated information, assessed by PCI. Once this analysis is tested as outlined in the grant, then PCI could also be used at different stages of development in the model (built for this thesis) to test the prediction that the capacity for information integration should increase with developmental changes such as synaptic refinement (Tononi, 2008, 2012).

Future research will also apply the theoretical notions developed in Aims #1 & 2 *in vivo*. This research program would be to find the spatiotemporal grain size in the brain that has maximally irreducible cause-effect power (Φ^{Max}), or, in other words, which grain contains the "differences that make [the most] difference." The irreducibility and cause-effect power at candidate grains could be calculated experimentally, as outlined in (Tononi et al., in prep). For example, the cause-effect repertoires of individual neurons in a particular state (such as bursting) could be approximated by inferring the probability distribution of past and future networks states. This could be done experimentally by recording the activity of a population of neurons (via a methodology like two-photon calcium imaging) while stimulating some of those neurons optogenetically. Irreducibility could be assessed by noising connections (equivalent to enforcing the probability of firing to maximum entropy) across a partition. Such an analysis could be carried out over individual neurons (taken as the micro) and then over larger and larger coarse-grains. As Aim #2 demonstrated, as the spatiotemporal scale with Φ^{Max} is approached Φ tends to increase, which may allow assessing causal emergence in the real brain, despite technical challenges.

References

Casali AG, Gosseries O, Rosanova M, Boly M, Sarasso S, Casali KR, Casarotto S, Bruno M-A, Laureys S, Tononi G, Massimini M. A theoretically based index of consciousness independent of sensory processing and behavior. *Sci Transl Med* 5: 198ra105, 2013.

Esser SK, Hill S, Tononi G. Breakdown of effective connectivity during slow wave sleep: investigating the mechanism underlying a cortical gate using large-scale modeling. *J Neurophysiol* 102: 2096–111, 2009.

Esser SK, Hill SL, Tononi G. Modeling the effects of transcranial magnetic stimulation on cortical circuits. *J Neurophysiol* 94: 622–39, 2005.

Esser SK, Hill SL, Tononi G. Sleep homeostasis and cortical synchronization: I. Modeling the effects of synaptic strength on sleep slow waves. *Sleep* 30: 1617–30, 2007.

Hill S, Tononi G. Modeling sleep and wakefulness in the thalamocortical system. *J Neurophysiol* 93: 1671–98, 2005.

Hoel EP, Marshall W, Albantakis L, Tononi G. Can macro beat micro? Integrated information across spatiotemporal scales. In prep.

Hoel EP, Albantakis L, Tononi G. Quantifying causal emergence shows that macro can beat micro. *Proc Natl Acad Sci U S A* 110: 19790–5, 2013.

Kandel ER, Markram H, Matthews PM, Yuste R, Koch C. Neuroscience thinks big (and collaboratively). *Nat Rev Neurosci* 14: 659–64, 2013.

Marshall W, Albantakis L, Tononi G. Blackboxing: A causal approach to understanding the physical world. In prep.

Mayer-Schönberger V, Cukier K. *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt, 2013.

Oizumi M, Albantakis L, Tononi G. From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Comput Biol* 10: e1003588, 2014.

Olcese U, Esser SK, Tononi G. Sleep and synaptic renormalization: a computational study. *J Neurophysiol* 104: 3476–93, 2010.

Tononi G, Boly M, Massimini M, Koch C. Integrated information theory: from consciousness to its physical substrate. In prep.

Tononi G. Consciousness as integrated information: a provisional manifesto. *Biol Bull* 215: 216–242, 2008.

Tononi G. Integrated Information Theory of Consciousness: An Updated Account. *Arch Ital Biol* 150: 56–90, 2012.

Appendix I

Supplementary Information

for

Quantifying causal emergence shows that macro can beat micro

Erik P Hoel[1], Larissa Albantakis[1], Giulio Tononi[1]
[1] Department of Psychiatry, University of Wisconsin, Madison, WI, USA

Published in:
The Proceedings of the National Academy of Sciences 2013; 110(49): 19790-19795

(S1) Effect coefficient and effectiveness (Eff) expressed as determinism and degeneracy

The state-dependent $\text{coefficient}(s_0) = \frac{\text{effect information}(s_0)}{\log_2(n)}$ can be described as a function of two $\log_2$ (n) terms, the determinism and degeneracy coefficient. To derive these two terms, the effect information(s_0), the distance between the effect repertoire ($S_F|s_0$) and the unconstrained repertoire of effects U^E, is split into the distance between ($S_F|s_0$) and the uniform distribution U with $p(s_U) = 1/n$, and a residual term:

$$\begin{aligned}
\text{Effect Information}(s_0) &= D_{\text{KL}}((S_F|s_0), U^E) = \sum_{s_F \in U^E} p(s_F|s_0) \log_2 \left(\frac{p(s_F|s_0)}{p(s_F)} \right) && (1)\\
&= \sum_{s_F \in U^E} p(s_F|s_0) \log_2 \left(\frac{p(s_F|s_0)}{p(s_U)} + \frac{p(s_U)}{p(s_F)} \right) && (2)\\
&= \sum_{s_F \in U^E} p(s_F|s_0) \left(\log_2 \left(\frac{p(s_F|s_0)}{p(s_U)} \right) - \log_2 \left(\frac{p(s_F)}{p(s_U)} \right) \right) && (3)\\
&= \sum_{s_F \in U^E} p(s_F|s_0) \log_2 \left(\frac{p(s_F|s_0)}{p(s_U)} \right) - \sum_{s_F \in U^E} p(s_F|s_0) \log_2 \left(\frac{p(s_F)}{p(s_U)} \right) && (4)\\
(\text{using } p(s_U) = 1/n) &= \sum_{s_F \in U^E} p(s_F|s_0) \log_2 \left(n \cdot p(s_F|s_0) \right) - \sum_{s_F \in U^E} p(s_F|s_0) \log_2 \left(n \cdot p(s_F) \right) && (5)\\
&= D_{\text{KL}}((S_F|s_0), U) - \sum_{s_F \in U^E} p(s_F|s_0) \log_2 \left(n \cdot p(s_F) \right), && (6)
\end{aligned}$$

where s_F denotes a state of the system S_F at t+1 with probabilty $p(s_F)$ according to the unconstrained distribution of effects U^E. s_0 is the present system state. The determinism coefficient is then the left term in line 5 and 6:

$$\text{Determinism coefficient}(s_0) = \frac{\sum_{s_F \in U^E} p(s_F|s_0) \log_2 \left(n \cdot p(s_F|s_0) \right)}{\log_2(n)} = \frac{D_{\text{KL}}((S_F|s_0), U)}{\log_2(n)}, \qquad (7)$$

the degeneracy coefficient the right term:

$$\text{Degeneracy coefficient}(s_0) = \frac{\sum_{s_F \in U^E} p(s_F|s_0) \log_2 \left(n \cdot p(s_F)\right)}{\log_2(n)}, \tag{8}$$

as defined in the main article.

The effectiveness (Eff) of a system assesses the causal relations in a system in a state-independent

$$\text{Eff}(S) = \frac{\text{EI}(S)}{\log_2(n)} = \frac{\langle \text{Effect Information}(s_0) \rangle}{\log_2(n)} = \frac{\sum_{s_0 \in U^C} p(s_0) D_{\text{KL}}((S_F|s_0), U^E)}{\log_2(n)}, \tag{9}$$

where the effective information EI(S) is the average effect information of all system states s_0, distributed according to U^C, the unconstrained repertoire of causes, which is identical to the uniform distribution U, thus here $p(s_0) = 1/n$. EI(S) can then be divided in the same way as the state-dependent effect information:

$$\begin{aligned}
\text{EI}(S) &= \langle \text{Effect Information}(s_0) \rangle & (10)\\
&= \langle D_{\text{KL}}((S_F|s_0), U) - \sum_{s_F \in U^E} p(s_F|s_0) \log_2 \left(\frac{p(s_F)}{p(s_U)}\right) \rangle & (11)\\
&= \langle D_{\text{KL}}((S_F|s_0), U) \rangle - \langle \sum_{s_F \in U^E} p(s_F|s_0) \log_2 \left(\frac{p(s_F)}{p(s_U)}\right) \rangle & (12)\\
&= \langle D_{\text{KL}}((S_F|s_0), U) \rangle - \sum_{s_0 \in U^C} p(s_0) \sum_{s_F \in U^E} p(s_F|s_0) \log_2 \left(\frac{p(s_F)}{p(s_U)}\right) & (13)\\
&= \langle D_{\text{KL}}((S_F|s_0), U) \rangle - \sum_{s_F \in U^E} p(s_F) \log_2 \left(\frac{p(s_F)}{p(s_U)}\right) & (14)\\
&= \langle D_{\text{KL}}((S_F|s_0), U) \rangle - D_{\text{KL}}(U^E, U). & (15)
\end{aligned}$$

The last equality is due to the fact that $p(s_F)$ is the probability of state s_F to occur at t_{+1} following U^E, the unconstrained distribution of effects (future states) obtained by setting the system S at t_0 into all possible states s_0 with equal probability $p(s_0) = 1/n$.

Both, indeterminism and degeneracy at the micro level may be indicative of causal emergence (see Discussion, main text). Note that in previous work, it was suggested that a convergence of two causes onto the same effect - an instance of degeneracy - may actually disqualify the micro level from causation (List and Menzies, 2009; Yablo, 1992) (though see (Shapiro and Sober, 2012)).

(S2) Effective information EI(S) expressed in terms of cause and effect information and mutual information MI

The effective information of a system, EI(S), can be obtained as the expected value of the cause or effect information. Moreover, EI(S) is identical to the mutual information $MI(U^C;U^E)$: the MI between the system S set to all possible counterfactuals (system states) with equal probability (unconstrained repertoire of causes, U^C) and the resulting distribution of system states at the next timestep (unconstrained repertoire of effects, U^E). Note that EI was originally introduced as a measure of causal influence of one subset of a system over another (Tononi and Sporns, 2003), while here it captures the overall

effectiveness of system S onto itself (see (Ay and Polani, 2008) and (Korb et al. 2011) for related measures).

In the following derivation, we start from the definition of EI(S) as the average effect information of all system states s_0 as counterfactual causes (distributed according to U^C with equal probability $p(s_0) = 1/n$ for all system states):

$$\mathrm{EI}(S) = \langle \text{Effect Information}(s_0) \rangle = \sum_{s_0 \in U^C} p(s_0) D_{\mathrm{KL}}((S_F|s_0), U^E) = \quad (1)$$

$$(\text{using } p(s_0) = 1/n \; \forall s_0) = \frac{1}{n} \sum_{s_0 \in U^C} D_{\mathrm{KL}}((S_F|s_0), U^E). \quad (2)$$

Using Bayes' Rule and time invariance we then show that the average effect information is indeed equivalent to the mutual information $MI(U^C;U^E)$ and to the expected value of the cause information, which is the average cause information of each accessible state at t_0, weighted by $p(s_0)$ according to U^E:

$$\mathrm{EI}(S) = \langle \text{Effect Information}(s_0) \rangle = \mathrm{MI}(U^C; U^E) = \langle \text{Cause Information}(s_0) \rangle. \quad (3)$$

In detail:

$$\begin{aligned}
\mathrm{EI}(S) &= \langle \text{Effect Information}(s_0) \rangle = \sum_{s_0 \in U^C} p(s_0) D_{\mathrm{KL}}((S_F|s_0), U^E) = && (4)\\
&= \sum_{s_0 \in U^C} p(s_0) \sum_{s_F \in U^E} p(s_F|s_0) \log_2 \left(\frac{p(s_F|s_0)}{p(s_F)} \right) = && (5)\\
&= \sum_{s_0 \in U^C} \sum_{s_F \in S_F} p(s_0) p(s_F|s_0) \log_2 \left(\frac{p(s_F|s_0)}{p(s_F)} \right) = && (6)\\
\text{(Bayes' Rule)} &= \sum_{s_0 \in U^C} \sum_{s_F \in U^E} p(s_0, s_F) \log_2 \left(\frac{p(s_0, s_F)}{p(s_0) p(s_F)} \right) = && (7)\\
&= \mathrm{MI}(U^C; U^E) = && (8)\\
\text{(time invariance)} &= \sum_{s_P \in U^C} \sum_{s_0 \in U^E} p(s_P, s_0) \log_2 \left(\frac{p(s_P, s_0)}{p(s_P) p(s_0)} \right) = && (9)\\
\text{(Bayes' Rule)} &= \sum_{s_P \in U^C} \sum_{s_0 \in U^E} p(s_0) p(s_P|s_0) \log_2 \left(\frac{p(s_P|s_0)}{p(s_P)} \right) = && (10)\\
&= \sum_{s_0 \in U^E} p(s_0) \sum_{s_P \in U^C} p(s_P|s_0) \log_2 \left(\frac{p(s_P|s_0)}{p(s_P)} \right) = && (11)\\
&= \sum_{s_0 \in U^E} p(s_0) D_{\mathrm{KL}}((S_P|s_0), U^C) = \langle \text{Cause Information}(s_0) \rangle. && (12)
\end{aligned}$$

Mutual information is originally a statistical measure of how much information is shared between a source and a target (Cover and Thomas, 2006). In the present context, MI is applied between two time steps of a system that is first perturbed into all counterfactuals (alternative states) with equal probability and then observed at the next time step. Because of the system perturbations, MI here is a causal measure. In other words, EI(S) is the MI between the set of all possible causes U^C and the set of all their effects U^E. Usually, however, mutual information is calculated for observed distributions of system states and thus not a causal measure, but a statistical measure of correlation.

(S3) Bounds of cause and effect coefficients and effectiveness Eff(S)

In the following, we will show that the cause and effect coefficients, as well as the effectiveness Eff(S),

are bounded between 0 and 1 ($\in [0 \ldots 1]$).

$$\text{Cause coefficient}(s_0) = \frac{\text{Cause information}(s_0)}{log_2(n)} = \frac{D_{\text{KL}}((S_P|s_0), U^C)}{log_2(n)} \quad (1)$$

$$\text{Effect coefficient}(s_0) = \frac{\text{Effect information}(s_0)}{log_2(n)} = \frac{D_{\text{KL}}((S_F|s_0), U^E)}{log_2(n)} \quad (2)$$

$$\text{Eff}(S) = \frac{EI(S)}{log_2(n)} = \frac{\frac{1}{n}\sum_{s_0 \in U^C} D_{\text{KL}}((S_F|s_0), U^E)}{log_2(n)} = \langle \text{Effect coefficient}(s_0) \rangle \quad (3)$$

The lower bound (0) is given by the fact that the Kullback-Leibler Divergence (D_{KL}) is always non- negative (Gibbs' inequality). Since the cause and effect information are expressed in terms of D_{KL} and the state independent effective information EI(S) is just an average of the state-dependent values, neither of the three coefficients can be negative. It thus remains to show that cause and effect coefficients cannot exceed 1.

The cause information(s_0) is the D_{KL} between the cause repertoire ($S_P|s_0$) and U^C, the unconstrained cause repertoire, which is identical to the uniform distribution with $p(s_P) = 1/n \; \forall s_P$. It follows that

$$\text{Cause information}(s_0) = D_{\text{KL}}((S_P|s_0), U^C) = \sum_{s_P \in U^C} p(s_P|s_0) \log_2 \left(\frac{p(s_P|s_0)}{p(s_P)} \right) = \quad (4)$$

$$= \sum_{s_P \in U^C} p(s_P|s_0) \log_2 \left(n \cdot p(s_P|s_0) \right) \quad (5)$$

$$(\text{since } p(s_P|s_0) \leq 1) \leq \sum_{s_P \in U^C} p(s_P|s_0) \log_2(n) = \log_2(n), \quad (6)$$

and thus

$$\text{Cause coefficient}(s_0) \leq 1. \quad (7)$$

The effect information(s_0) is the D_{KL} between the effect repertoire ($S_F|s_0$) and U^E, the unconstrained

$$p(s_F) = \sum_{s_0 \in U^C} p(s_F|s_0) \cdot p(s_0), \tag{8}$$

where $p(s_0) = 1/n \; \forall \, s_0$ and thus:

$$p(s_F|s_0) \leq n \cdot p(s_F), \; \forall s_F. \tag{9}$$

Using eq. 9, if follows that:

$$\begin{aligned} \text{Effect information}(s_0) &= D_{\mathrm{KL}}((S_F|s_0), U^E) = \sum_{s_F \in U^E} p(s_F|s_0) \log_2 \left(\frac{p(s_F|s_0)}{p(s_F)} \right) = & (10) \\ (\text{eq.9}) &\leq \sum_{s_F \in U^E} p(s_F|s_0) \log_2 \left(\frac{n \cdot p(s_F)}{p(s_F)} \right) = \sum_{s_F \in U^E} p(s_F|s_0) \log_2(n) & (11) \\ &= \log_2(n), & (12) \end{aligned}$$

and thus

$$\text{Effect coefficient}(s_0) \leq 1. \tag{13}$$

Finally, since the effect coefficient(s_0) $\in [0 \ldots 1] \; \forall \, s_0$, also its average over all system states, the state independent effectiveness Eff(S) $\in [0 \ldots 1]$.

(S4) Causal reduction

To complement the examples of causal emergence in the main text, we here provide an example in which causal reduction is called for. In Fig. AI.1, a macro mechanism works

as an XOR logic gate (as an isolated part of a larger circuit board) with inputs X,Y and output Z (Fig. AI.1A). At the macro level, the system (XOR,X,Y,Z) generates 2 bits of EI over 1 macro time step T_X (the XOR operates after a 'decision' period where it processes the input) and $Eff(S_M)=0.5$. The macro XOR gate is actually composed of (supervenes upon) 9 deterministic micro logic gates (COPY, NOT, AND, OR). In this case, however, causal interactions are stronger at the micro level and over a single micro time step t_X ($EI(S_M)=7.43$ bits and $Eff(S_M)=0.83$). Thus, CE=-5.43 bits, corresponding to negative causal emergence, i.e. reduction. Note that in this case the micro circuit is deterministic and minimally degenerate (0.17), so the macro cannot offset the loss of effective information due to its reduced size by a gain in determinism or a reduction in degeneracy.

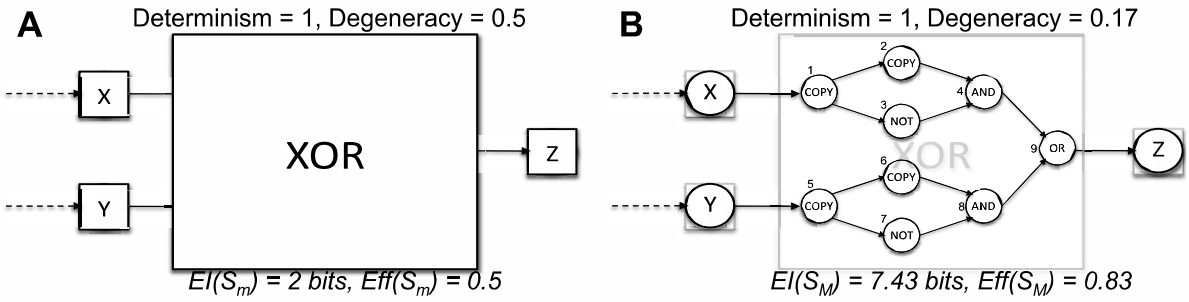


Fig. AI.1. Causal reduction. (A) A part of a larger circuit is presented, which performs a macro XOR logic function over its inputs X, Y, and outputs to X. (B) At the micro level, the XOR consists of 9 deterministic logic gates. The system is deterministic at both the macro and micro level. Moreover, the degeneracy coefficient at the micro level is lower than at the macro level. Therefore in this case, the micro beats the macro, leading to causal reduction. **CE(S) = 5.43 bits**.

Note that, in order to demonstrate this case of causal reduction, we have assumed that a deterministic micro circuit underlies the above macro circuit. In general, however, real digital circuits are often built from many stochastic analog micro elements in a

highly degenerate manner, to compensate for noise at the lower level and to create deterministic macro elements. In this way, digital circuits and other engineered systems follow similar design principles as the more physiological examples presented in the main text. Consequently, there is the potential for either causal emergence or reduction in digital circuits, depending on the underlying micro level, just as in physiological systems.

More generally, the notion of causal reduction ($CE < 0$) stands in contrast to previous accounts of reduction that focused on the relationship between scientific theories and whether or not they are reducible to one another (Nagel, 1961). In the present account based on causal analysis, the focus is instead on the relationship between micro and macro levels of mechanisms. This account reveals why there is a bias in favor of reductionism in mechanistic scientific explanations. The bias is understandable given that, everything else being equal, the micro would always beat the macro: being more detailed by definition, the micro has an inherent advantage in how informative its causal mechanisms are. This inherent advantage is captured quantitatively in causal analysis because the micro can benefit from both ΔI_{Eff} and ΔI_{Size}, whereas the macro can only gain from ΔI_{Eff}.

(S5) Causal emergence in a system with causally heterogeneous elements

While the examples in the main text (with the exception of Fig.6) all have macro elements with underlying unconnected and causally equivalent micro elements, this is not a necessity for causal emergence. In Fig. AI.2A, the 6 micro elements are fully interconnected and causally heterogeneous. The elements are structured into 2 groups

{ABC, DEF} due to different intra- and inter-group mechanisms: within each group, if the sum of intra-group connections = 0, all elements stay 0 (inactive) the next time step. However, if the sum of inter-group connections = 3 (synchronous activity from the other group), all elements turn 1, unless they are all 0, in which case they become spontaneously active (1) with probabilities: p(A/D)=0.45; p(B/E)=0.5; p(C/F)=0.55. Since the micro TPM is noisy, EI(S_m) = 1.13 bits and Eff(S_m) = 0.19 (Fig. AI.2B). The optimal macro grouping S_M (Fig. AI.2C) has a more deterministic TPM (Fig. AI.2D), EI(S_M) = 1.84 bits and Eff(S_M) = 0.58. Thus, the macro supersedes the micro (CE(S) = 0.72 bits) despite its reduced repertoire size, because it counteracts noise by responding almost deterministically to synchronous activity over inter-group connections.

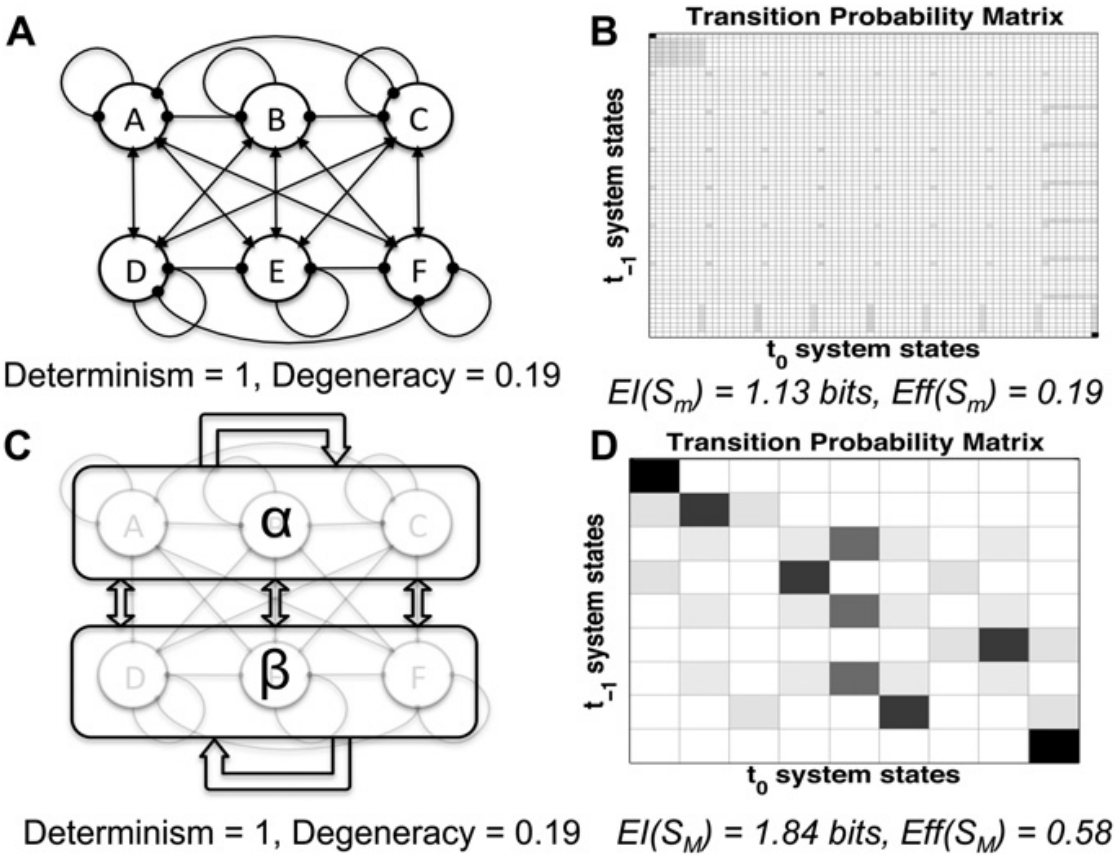


Fig. AI.2. Causal emergence in a system with differentiated connectivity. (A) Micro system with six elements. Regular and rounded arrows indicate intergroup and intragroup connecitons, respectively. (B) Noisy micro-level TPM. (C) Macro system where each macro element receives inputs from itself and the other macro element. (D) More deterministic macro-level TPM. **CE(S) = 0.72 bits**.

The neural-like system of Fig. 6 in the main text has equivalent spatial properties to the example system of Fig. AI.2 (fully connected, causally heterogeneous elements, sensitive to differences in intra- and inter-connections). In addition, it has the same temporal properties as the system shown in Fig. 5 (main text), with 2^{nd}-order Markov mechanisms at the micro level. The system's states space at the micro level thus contains 2^{18} states, which prohibited an exhaustive search for the optimal macro level. Nevertheless, the spatiotemporally emergent macro grouping shown in Fig.6B (main text) is assumed to be the optimal macro grouping based on the results obtained from the examples of Fig. AI.2 and Fig.5 (main text).

(S6) Applicability - Network motifs as indicators of emergence

Measuring EI exhaustively, across all micro/macro levels, is not feasible for large systems. This is because, assuming N binary elements, $B_N - 1$ (N^{th} Bell number) possible groupings of those micro elements into macro elements exist, each of which entails

$$\prod_{j=1}^{k} \left(B_{m(j)+1} - 1\right)$$

possible groupings of micro into macro states, where k is the number of macro elements with m(j) micro elements each. The number of EI computations to determine the spatio-

temporal grain with maximal EI thus increases dramatically with N (N = 1: 1 N = 2: 5, N = 3: 27, N = 4: 180 computations, etc.) if calculated exhaustively.

In large, complex networks where an exhaustive causal analysis is unfeasible, overrepresented network motifs could already indicate whether the network as a whole is biased towards emergence or reduction. For example, the two most common network motifs shared by the gene networks in E. Coli and the brain of C. Elegans are the feed-forward loop and the bi-fan (Milo et al., 2002). Both these network motifs mimic in their connectivity precisely the micro element groups that made up the optimal (winning) macro elements in our chosen examples. In Fig.2 (main text), the first spatial example, the macro elements are bi-fans, while in Fig.6 (main text), the first temporal example, the macro elements are feed-forward loops. These are perhaps the simplest possible functionally relevant macro elements. Both, the bi-fan and the feedforward loop show causal convergence (degeneracy) in either space or time. A greater than random prevalence of these or similar network motifs, paired with some amount of intrinsic noise in the system, may indicate that the system operates at a macro level.

References

Ay N, Polani D. Information flows in causal networks. *Adv Complex Syst* 11:1741, 2008.

Cover TM, Thomas JA. *Elements of information theory. Wiley-interscience*. 2006.

Korb KB, Nyberg EP, Hope L. *Causality in the Sciences*, eds Illari P, Russo F, Williamson J (Oxford University Press, Oxford), p.628-652, 2011.

List C, Menzies P. Non-reductive physicalism and the limits of the exclusion principle. *Journal of Philosophy* CVI(9): 475-502., 2009.

Milo R, Shen-Orr S, Itzkovitz S. Network motifs: simple building blocks of complex networks. *Science* 298:824-827, 2002.

Nagel E. The structure of science. *American Journal of Physics* 29:716-716, 1961.

Shapiro L, Sober E. Against proportionality. *Analysis* 72:89-93, 2012.

Tononi G, Sporns O. Measuring information integration. *BMC Neurosci* 4:31, 2003.

Yablo S. Mental causation. *Philos Rev* 101:245-280., 1992.

Appendix II

Supplementary Information

for

Information integration and the spatiotemporal scale of consciousness

Erik P Hoel[1], Larissa Albantakis[1], William Marshall[1], Giulio Tononi[1]
[1] Department of Psychiatry, University of Wisconsin, Madison, WI, USA

In preparation.

Irreducible selectivity and shift in selectivity

The causal properties of irreducible selectivity and selectivity-shift can be further understood from a geometric perspective. From this perspective, cause-effect repertoires are points in a high-dimensional metric space, wherein the EMD is the metric between repertoires. The irreducible selectivity and selectivity-shift quantities summarize the relationship between three points in this metric space: the unpartitioned repertoire (UP), the partitioned repertoire (P), and a maximum entropy distribution (H).

The irreducible selectivity of a concept measures the increase or decrease in repertoire entropy as a result of the *MIP*. Geometrically, this captures whether the partitioned repertoire is closer (selectivity > 0) or further (selectivity < 0) from maximum entropy than the unpartitioned distribution.

The selectivity-shift of a concept captures the degree to which the partitioned repertoire selects different system states than the unpartitioned distribution. This is any distance in the high dimensional metric space between the partitioned and unpartitioned distributions which is not in the direction of maximum entropy. Shift can be captured by the increase in distance travelled when moving from the unpartitioned distribution to the maximum entropy distribution by way of the partitioned distribution, as opposed to taking the most direct route.

Different arrangements of points (cause-effect repertoires) in a two dimensional metric space are shown in Figure AII.1. We investigate the irreducible selectivity and selectivity-shift values for each arrangement. The unpartitioned repertoire, in its distance from the maximum entropy distribution, defines a sphere of points equidistant from the maximum entropy distribution, such that if the partitioned repertoire is on the sphere then the irreducible selectivity is zero. When the partitioned repertoire lies inside the contour then the difference in selectivity is positive, and if it is outside the contour line it is negative. If the partitioned repertoire lies on a line between the unpartitioned repertoire and maximum entropy distribution then selectivity-shift is zero, otherwise there is a positive value of selectivity-shift. Only when the partitioned repertoire is equivalent to the unpartitioned repertoire (the mechanism is reducible and the concept does not exist) is it possible that both selectivity and selectivity-shift are both zero.

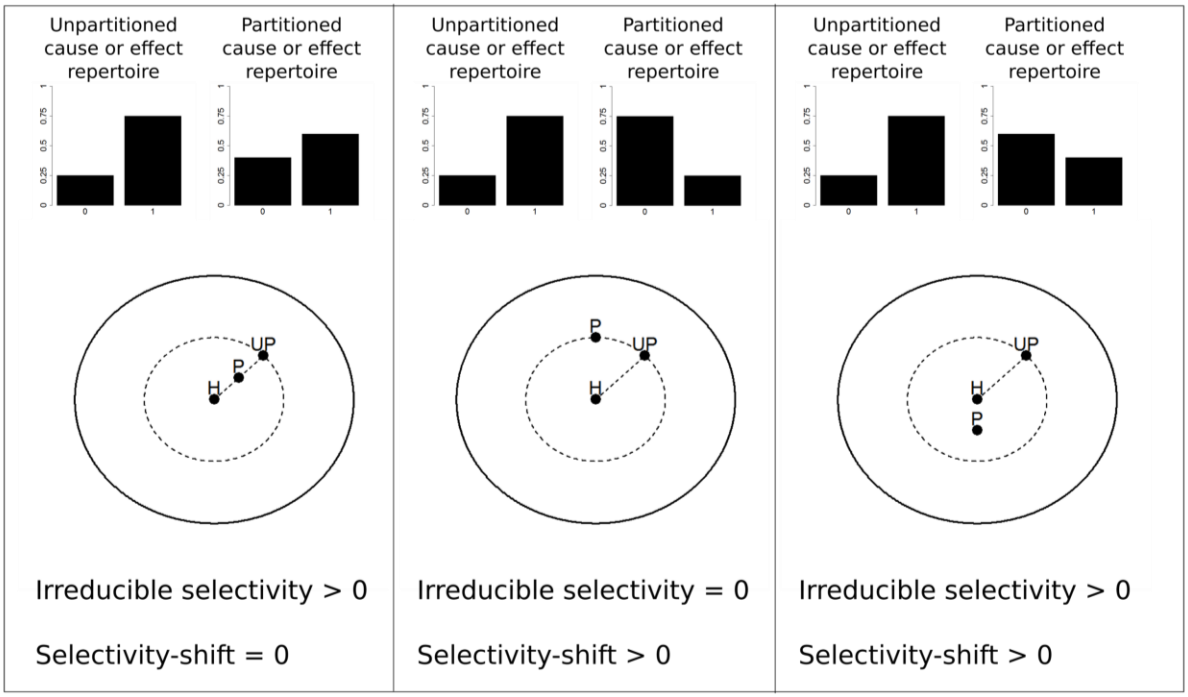


Figure AII.1: Relationship of irreducible selectivity and selectivity-shift to the position of unpartitioned repertoire, partitioned repertoire and maximum entropy distribution in metric space. Left: If the partitioned repertoire lies on the line between the unpartitioned and max entropy distribution, then the selectivity-shift is zero. Middle: If the unpartitioned and partitioned are equally distance from maximum entropy then selectivity is zero and selectivity-shift is positive. Right: If the partitioned repertoire lies within the inner circle, but not on a direct line between H and UP, then both irreducible selectivity and selectivity-shift are positive.

Relationship to φ

Here we show how the integrated information value of a concept (phi) can be

decomposed into three factors: purview size, irreducible selectivity, and selectivity-shift.

$$\varphi = EMD(S||S_{MIP})$$

$$= (EMD(S||S_{MIP}) + EMD(S_{MIP}||H) - EMD(S||H))$$

$$+ \ (EMD(S||H) - EMD(S_{MIP}||H))$$

$$= size* \ (irreducible \ selectivity + \ selectivity\text{-}shift)$$

We also show some additional properties of irreducible selectivity and selectivity-shift. First, integrated information is bounded by size (Marshall et al. in prep), which in turn bounds the sum of irreducible selectivity and selectivity-shift:

$$\varphi \leq size$$

$$size * (irreducible\ selectivity + selectivity\text{-}shift) \leq size$$

$$irreducible\ selectivity + selectivity\text{-}shift \leq 1$$

As a partition can either decrease or increase the entropy, irreducible selectivity can be either positive or negative in value, bounded in a range of length 1 (-0.5, 0.5). This bound uses the fact that the maximum EMD from any distribution to the maximum entropy distribution is at most 1/2*size. (Marshall et al. in prep). To see the upper bound

$$\begin{aligned} irreducible\ selectivity &= (EMD(S \,||\, H) - EMD(S_{MIP} || \, H))/size \\ &= \left(\frac{size}{2} - 0\right)/size \\ &= 0.5 \end{aligned}$$

and to see the lower bound,

$$\begin{aligned} irreducible\ selectivity &= (EMD(S \,||\, H) - EMD(S_{MIP} || \, H))/size \\ &= \left(0 - \frac{size}{2}\right)/size \\ &= -0.5 \end{aligned}$$

The selectivity-shift value is bounded from (0, 1). It is non-negative; since we are working in a metric space this result follows from the triangle inequality. To see that the maximum selectivity-shift value is 1, first the maximum distance from the unpartitioned repertoire to the partitioned repertoire is bounded by

$$EMD(S \,||\, S_{MIP}) \leq EMD(S \,||\, H) + size \,/\, 2.$$

To see why this is true, consider a fixed reference point (unpartitioned distribution) in the circular space of repertoires (see Figure S1). The furthest point on the circle from this reference point is the boundary point that is in the direction of the center of the circle (maximum entropy distribution). Thus the largest distance possible is to take the distance to the center of the circle plus the distance from the center to the boundary (radius, size/2). Plugging this relationship into the formula for shift,

$$\begin{aligned} irreducible\ shift &\leq (EMD(S||S_{MIP}) + EMD(S_{MIP}||H) - EMD(S||H)) \,/\, size \\ &\leq EMD(S||H) + size \,/\, 2 + EMD(S_{MIP}||H) - EMD(S||H)) \,/\, size \\ &= (size \,/\, 2 + EMD(S_{MI\square}||H)) \,/\, size \\ &\leq (size \,/\, 2 + size \,/\, 2) \,/\, size \\ &= 1 \end{aligned}$$